Connected Networks in Principal Value-Added Models

Brendan Bartanen
Texas A&M University

Aliza N. Husain
University at Buffalo

A growing literature uses value-added (VA) models to quantify principals' contributions to improving student outcomes. Principal VA is typically estimated using a connected networks model that includes both principal and school fixed effects (FE) to isolate principal effectiveness from fixed school factors that principals cannot control. While conceptually appealing, high-dimensional FE regression models require sufficient variation to produce accurate VA estimates. Using simulation methods applied to administrative data from Tennessee and New York City, we show that limited mobility of principals among schools yields connected networks that are extremely sparse, where VA estimates are either highly localized or statistically unreliable. Employing a random effects shrinkage estimator, however, can alleviate estimation error to increase the reliability of principal VA.

VERSION: May 2021

Connected Networks in Principal Value-Added Models

Brendan Bartanen
Texas A&M University

Aliza N. Husain
University at Buffalo

May 6, 2021

Abstract

A growing literature uses value-added (VA) models to quantify principals' contributions to improving student outcomes. Principal VA is typically estimated using a connected networks model that includes both principal and school fixed effects (FE) to isolate principal effectiveness from fixed school factors that principals cannot control. While conceptually appealing, high-dimensional FE regression models require sufficient variation to produce accurate VA estimates. Using simulation methods applied to administrative data from Tennessee and New York City, we show that limited mobility of principals among schools yields connected networks that are extremely sparse, where VA estimates are either highly localized or statistically unreliable. Employing a random effects shrinkage estimator, however, can alleviate estimation error to increase the reliability of principal VA.

Author Note: Brendan Bartanen (bartanen@tamu.edu) is an assistant professor in the College of Education and Human Development at Texas A&M University. Aliza N. Husain is an assistant professor in the Graduate School of Education at the University at Buffalo. We are grateful to Dale Ballou, Jason Grissom, Matt Kraft, and Jim Wyckoff for their comments on the paper. Any errors are our own.

1 Introduction

A growing literature seeks to estimate *principal value-added* (VA): statistical models that isolate the contributions of individual principals to school performance, most often conceptualized as student test score gains, in an education production function. VA methods applied to principals can provide answers to two important questions: (1) How important are principals as inputs to student learning? (2) Who is an effective principal? Extant work consistently finds that principals matter, with the magnitude of principal effects typically ranging between 0.05 and 0.20 student-level standard deviations (SD) (Grissom, Kalogrides, and Loeb 2015; Bartanen 2020; Dhuey and Smith 2018; Chiang, Lipscomb, and Gill 2016; Branch, Hanushek, and Rivkin 2012). In other words, a 1 SD increase in principal VA increases student achievement by 0.05 to 0.20 SD.

A key empirical challenge to estimating principal effects is to account for myriad school- or district-level factors that affect student learning but that principals cannot control. For example, principals cannot control the neighborhood in which the school is located and they similarly face constraints around teacher hiring and/or retention due to district policies or local labor market conditions. These school factors are often difficult to measure and may not be well-captured by available proxy measures, such as average student demographics. The approach taken in prior studies is to estimate a two-way regression model that includes principal and school fixed effects (FE), under the assumption that such unobserved school factors are largely fixed across time. In this model, identification of the principal fixed effects (i.e., the VA estimates) is restricted to within-school variation, but additional across-school comparisons of principals are possible if some principals work in multiple schools over time. Mobility groups formed by principals and schools due to principals transitioning across schools are termed "connected networks" in the principal VA literature (e.g., Bartanen 2020; Chiang, Lipscomb, and Gill 2016; Burkhauser 2017).[1]

[1] While defined later, a "connected network" comprises the largest possible set of schools in which every school has had at least one principal move to at least one other school in the network.

The appeal of the connected networks model is that it can yield VA estimates where principals have a large comparison set, while also avoiding misattribution of school effects to principals. Despite its common use in estimating principal VA, however, our understanding of the properties of estimates from the connected networks approach remains limited. The inclusion of school FE creates additional challenges stemming from the limited mobility of principals among schools. Whereas other applications of two-way network models may benefit from observing many individuals working in multiple firms (or teachers in multiple schools), a school typically has only one principal at time and a majority of principals lead only one school in their career. This limited mobility potentially leads to weakly identified FE estimates that are unreliable measures of principal quality and overstate the magnitude of principals' effects (Jochmans and Weidner 2019).

Connected networks models further hinge on the fundamental assumption that a principal's effectiveness is the same in any two schools. This assumption allows for the indirect comparison of principals who never worked in the same school, which is a key practical benefit of the connected networks model. Given prior work demonstrating that leadership is a relational process and that principals' impacts on student achievement are largely mediated through other school-level factors (e.g., Hallinger and Heck 1998; Sebastian and Allensworth 2012), such an assumption may be unrealistic. In particular, there may exist principal–school complementarities whereby part of a principal's effect reflects how well matched they are to a particular school context (Dhuey and Smith 2018).

Using a simulation approach, this paper helps to fill a gap by investigating the accuracy of connected network models for estimating principal effects. Simulation studies have the advantage of creating controlled conditions where the performance of VA estimators can be tested according to different assumptions about the data-generation process. In particular, this approach allows us to focus on the aforementioned issues related to the connected networks approach, as opposed to principal VA modeling more generally. In that sense, our simulation is conceptually similar to those used to examine the accuracy of teacher VA models under different assumptions about the nonrandom sorting of students to teachers (e.g., Guarino, Reckase, and Wooldridge 2015; Guarino et al. 2015). To supplement the insights drawn from the simulation, we also provide a brief empirical application using actual test scores.

Our simulation is built from administrative datasets from Tennessee and New York City (NYC). That is, using the connected networks formed by actual mobility patterns of principals and schools over long panels, we generate simulated test scores where the true principal effects are known. We then apply VA models to the simulated data and compare the estimated and true principal effects. We consider two questions regarding the performance of principal VAMs: (1) How accurately do VAMs rank principals according to their true effects? (2) To what extent does the magnitude of principal VA accurately reflect the true magnitude of principals' effects? Answers to these questions can provide insight about whether principal VA models are likely to provide accurate results in real-world conditions.

Our results uncover a key tradeoff between the statistical precision of principal VA estimates and their practical utility. Even in large-scale datasets where we observe thousands of principal transitions, the underlying network structure of principals and schools is very weak. This yields two distinct types of connected networks. First, many principals belong to small networks that contain one or two schools, meaning that their estimated VA reflects performance relative to only a handful of other principals. While principal VA from small connected networks is precisely estimated, such localized performance measures may lack practical usefulness (e.g., as an accountability metric). Principals in large networks, on the other hand, can be compared to hundreds of other principals. However, the underlying network is weakly connected, undermining the reliability of VA estimates and producing inaccurate rankings of principals. Further, because schools in large connected networks are typically linked through only one or two mobile principals, inaccuracy is amplified substantially in the presence of principal–school complementarities.

A similar tradeoff exists for using principal VA models to understand the magnitude of principals' impacts on student outcomes. In small networks, school FE erroneously eliminate part of the real difference in principal quality, leading to an understatement of the importance of principals. While large networks circumvent this problem, they *overstate* the magnitude of principal effects because of the estimation error introduced by weak network structures.

Given the estimation error of principal VA estimates in large networks, we further examine whether shrinkage approaches can improve correlations between principals' estimated and true effects. Employing a mixed model that treats principal effects as random greatly reduces estimation error in large networks. This method improves the precision of principal VA and yields a substantially more accurate estimate of the magnitude of principal effects.

This paper contributes to our understanding of principal VA models, where evidence on their validity and reliability remains limited. This dearth of evidence stands in stark contrast to the teacher VA literature, where a large number of studies have investigated these properties (see Koedel, Mihaly, and Rockoff 2015, for a review). We also contribute to a larger literature utilizing two-way regression models in the context of matched employer–employee datasets. Most notably, we provide an application of recent theoretical work (e.g., Jochmans and Weidner 2019; Verdier 2018; Kline, Saggio, and Sølvsten 2018) concerning inference of fixed effects estimated from network data. We also build on related applications using school fixed effects to control for unobserved school heterogeneity in value-added modeling, including for teachers (Mansfield 2015) and teacher preparation programs (e.g., Mihaly et al. 2013). These applications highlight some key challenges for analyzing connected networks that we consider in the context of the principal labor market, with the added benefit of a simulation analysis that can provide deeper insight around the accuracy of VA models with school FE.

The rest of the paper is organized as follows. Section 2 outlines the estimation of principal effects in the connected networks model and the challenges it creates for the accuracy of principal VA. In section 3 we describe our simulation procedures and the data sources that form the network structure for our simulation. Section 4 presents our baseline simulation results, results using shrinkage approaches, and an empirical application that uses actual test scores from Tennessee. Section 5 concludes with a discussion of implications and areas for future work.

2 Connected Networks in Principal Value-Added Models

In this section, we first outline the basic empirical challenge of estimating principal effects on student outcomes, which is to separate principals' contributions from school factors they cannot control. This leads to the concept of *connected networks* of principals and schools. We then describe the challenges that connected network models introduce, which inform our simulation analysis.

2.1 The Principal and School Fixed Effects Model

Principal VA models seek to estimate the contribution of individual principals to student test score growth. As with any VA approach, a key empirical challenge is to isolate principal effectiveness from other factors that affect student outcomes but that are not attributable to principals. This is particularly challenging for principals, whose impacts on student outcomes are largely indirect. Whereas teachers provide direct instruction to students, principals affect student achievement growth through malleable school factors, such as establishing a positive school climate or hiring and retaining effective teachers (e.g., Grissom and Bartanen 2019; Kraft, Marinell, and Yee 2016; Jacob 2011; Hallinger and Heck 1998; Sebastian and Allensworth 2012; Coelli and Green 2012). School performance, however, is also affected by fixed school factors that principals cannot control, such as the neighborhood or the quality of school facilities (Chiang, Lipscomb, and Gill 2016; Grissom, Kalogrides, and Loeb 2015). Further, there may exist a correlation between principal quality and such factors, whereby certain schools are systematically more or less likely to be led by high-quality principals

(Branch, Hanushek, and Rivkin 2012; Grissom, Bartanen, and Mitani 2019).

On a fundamental level, a principal's VA estimate is the mean test score residual across all of the students who were in the school during the principal's tenure. A high-VA principal, then, is one whose students tend to have positive test score residuals. Crucial to the approach is that the residualization adequately accounts for school factors that should not be attributed to principal quality. Particularly necessary is adjusting for students' background characteristics to address the non-random sorting of students to schools. These controls typically include students' prior-year test scores, demographic characteristics (e.g., race/ethnicity, free/reduced-price lunch eligibility) and academic characteristics (e.g., special education status). VA models also commonly control for school-level averages of student background characteristics to further account for between-school heterogeneity.

Even with controls for observable student and school characteristics, there likely remains substantial unexplained school-level variation that is not part of a principal's effect. The quality of the school's neighborhood, for instance, may only be partially captured by students' background characteristics.[2] To better account for unmeasured school factors, prior studies estimating principal VA nearly universally include school fixed effects in the control vector, leading to the following model:

$$Y_{ijst} = \lambda(f(Y_{i,t-1})) + \omega\mathbf{X}_{it} + \phi\mathbf{Z}_{st} + \delta_j + \gamma_s + \epsilon_{ijst} \tag{1}$$

where δ_j and γ_s are vectors of indicator variables (i.e., fixed effects) for principals and schools, respectively.

[2] We provide empirical support for this claim in Appendix Table C.1, which shows variance component estimates from mixed models in Tennessee and New York City using different sets of controls. Even after controlling for prior-year test scores, student characteristics, and school-by-year means of student characteristics, the school-level variance component remains roughly equal in magnitude to the principal variance component, underscoring the potential for bias in principal VA estimates that do not account for unobserved school heterogeneity.

2.2 Connected Networks

Estimating effects for both principal and school is challenging because each school has only one principal at a time, meaning that in cross-sectional data δ_j and γ_s are perfectly collinear. With panel data of sufficient length, however, schools will have multiple principals, which creates the necessary within-school variation required to estimate coefficients for δ_j. The interpretation of principal VA estimates between models with and without school fixed effects, however, can be very different. Without school fixed effects, VA estimates produce a global ranking of all observed principals. When school FE are included, principal VA estimates produce *local* rankings. Specifically, with school FE, principals can only be compared within a *connected network* of schools, where a network is the largest possible set of schools in which every school has had at least one principal transfer to at least one other school in the network during the analysis period.

The size of a connected network can range from a single school to the entire set of schools, depending on the number of schools and years in the panel and the mobility patterns of principals across schools. A single-school network will result when none of the principals who worked in that school were observed working in a different school. A multi-school connected network will form when a principal who works in school A moves to school B. This connection allows for the comparison of all principals who ever worked in school A or school B. If school A or school B further has a principal who also worked in school C, the connected network grows to include all principals who ever worked in one of these three schools. Given sufficient mobility, a connected network can theoretically include the complete set of observed schools.

In practical applications, however, connected networks of principals and schools tend to be small, often comprised by only a single school. Even when longer panels of data are available, high principal attrition rates (i.e., exiting from the principalship entirely) mean that there are relatively few principals who transfer between schools. For instance, when examining data for Pennsylvania students in grades 4–8 from the 2008–09 to 2012–13 school years (a comparatively small panel), Chiang, Lipscomb, and

Gill (2016) find that 76 percent of connected networks are single-school and only 2.6% of networks included four or more schools. Similarly, of the connected networks present in Burkhauser (2017)'s data spanning 2005–06 to 2011–12, 57% were networks with only two principals, indicating that a majority of school leaders were being compared to only one other leader. In a statewide panel spanning 10 years, Bartanen (2020) observes almost 20% of principals in networks with 10 or more schools, with 39% of principals in single-school networks. In sum, connected networks of schools and principals generally tend to be small, limiting the comparison set of principals and leading to relatively local measures of principal quality.

While the interpretation of principal VA estimates in network models as local measures of principal effectiveness has been well-established in prior studies, there has been virtually no work that investigates their validity and reliability. Consequently, we know little about whether these models produce accurate measures of principals' effects on student outcomes. While the inclusion of school FE is conceptually appealing as a means to control for unobserved factors, it also introduces a number of additional challenges that have not been given much attention. Our analysis focuses on three of these challenges, which we outline below.

2.3 Estimation Error in Sparse Networks

While prior studies make clear that including school FE in principal VA models changes the interpretation of a principal's estimated effect, they fail to note that school FE introduces estimation error into each estimate, which lowers reliability and leads to upward bias in the estimated magnitude of principal effects. Recently, econometricians have paid increasing attention to two-way regression models using network data (Jochmans and Weidner 2019; Verdier 2018); the principal and school fixed effect model is one application of such models. When examining these models, Jochmans and Weidner (2019) demonstrate that the statistical precision of individual effects is determined by the connectivity structure of the underlying network. In the case of principal VA, limited mobility of principals among schools means both that many principals are in small networks and that estimates for principals in larger networks contain considerable noise. Intuitively, this noise comes from variance inflation, as indicator variables for each principal and school are highly correlated with one another.

Specifically, Jochmans and Weidner (2019) show that the two-way fixed effects model (in our case, principals and schools) can be analyzed as a weighted bipartite graph, where edges connecting principals to schools are weighted by the number student-by-year observations. The statistical precision of the principal fixed effects is determined by how strongly connected principals are within the given connected network. In particular, they demonstrate that bottlenecks in the network (i.e., where two larger sets of principals are connected only through a single principal) lead to variance inflation and, ultimately, imprecise VA estimates. Mathematically, these bottlenecks can be summarized by the smallest nonzero eigenvalue (λ_2) of the graph's normalized Laplacian matrix, where $\lambda_2 \to 0$ as the network becomes more sparse.

Variance inflation in principal VA models has not been formally investigated, though prior work tends to conclude that VA estimates are precise, given the large number of student-by-year observations that contribute to estimating each principal's effect. Thus, our analysis contributes by examining variance inflation due to the structure of connected networks, which we show theoretically using the techniques from Jochmans and Weidner (2019) and through our simulation that uses the actual connected networks of principals in Tennessee and New York City.

2.4 Downward Bias of the Variance in Small Networks

While estimation error due to sparse network structure creates an upward bias in the estimated magnitude of principal effects, an additional challenge introduced by the inclusion of school fixed effects is a *downward* bias in the estimated magnitude, concentrated in small networks that have few principals. To see this, consider a scenario where each school has two principals and no principals switch schools, thus making each school its own connected network. In expectation, the mean principal effect in each network is zero, but the observed mean will be nonzero due to sampling variation. In the connected networks approach, this sampling variation—which reflects real information about principal quality—will be captured by the school fixed effect, which creates a downward bias in the estimated variance of the principal effect that decreases as the number of principals increases. This downward bias works in the opposite direction as estimation error from the sparse network structure.

As connected networks grow larger, this downward bias in the variance will decrease, which will make the empirical distribution of VA estimates a better representation of the true variance of principal effects. As described in the previous section, however, large networks may suffer from greater estimation error. These dynamics may help to explain why prior studies reach somewhat different estimates of the magnitude of principal effects despite similar empirical approaches. Bartanen (2020) and Dhuey and Smith (2018), for instance, draw on long panels from statewide data (where more principals are in large networks) and find larger SD of principal VA estimates. By contrast, Grissom, Kalogrides, and Loeb (2015) and Branch, Hanushek, and Rivkin (2012) find lower SDs. In the former study, principal VA is estimated using just a single district across an 8-year panel. While Branch, Hanushek, and Rivkin (2012) draw on statewide data from Texas, they restrict the size of connected networks to a single school by estimating principal-by-school FE.

2.5 Principal–School Complementarities

As outlined in the connected networks section, principal VA models that include school FE result in VA estimates that are local to the principal's connected network. While the identifying variation is restricted to comparisons of principals who worked in the same school, indirect comparisons are made possible via principals who work in multiple schools across the study period. Intuitively, principal A, who worked in school X, can be compared to principal B, who worked in school Y, if there exists a principal C who worked in both schools. Connected networks can grow large as the length of the panel increases or as principals move between schools more frequently.

Underpinning these indirect comparisons of principals who never worked in the same school is the assumption that the effectiveness of the mobile principal (i.e., the one who connects the two schools) is fixed. More specifically, the connected networks approach requires an assumption that there are no complementarities or "match effects" between principals and schools. Taking the simple example above, comparing principal A in school X to principal B in school Y breaks down if principal C's effectiveness is different in school X versus school Y. If, for example, principal A and principal B are equally effective, but principal C is better matched in school X than school Y, then principal A will appear less effective than principal B in the connected networks model.

While there has been little empirical work examining principal–school complementarities in the context of VA models (see Dhuey and Smith 2018, for an exception), prior studies demonstrate that principals operate within the contexts of their schools. The idea of "fit" between an individual and their work environment was formally introduced in the industrial-organizational psychology literature, and coined "person-environment (PE) fit" (Kristof-Brown, Zimmerman, and Johnson 2005). PE fit refers to the compatibility between a person and their work environment, especially based on the match between their respective characteristics (Kristof-Brown, Zimmerman, and Johnson 2005; Miller et al. 2020). Direct examinations of this theory in education have been limited thus far (see Miller et al. 2020; Player et al. 2017; Jackson 2013, for some recent exceptions in the teacher literature); however recent work in the educational leadership literature has begun to explore related issues. For instance, research shows that Black principals increase the likelihood of Black teachers being hired and retained (Bartanen and Grissom 2021), and male teachers are more likely to turn over under female principals and request to transfer to schools with male principals (Husain, Matsa, and Miller 2018). Evidently, principals experience varied levels of success based on teacher demographics in their schools, speaking directly to the concept of PE fit.

At first blush, incorporating match effects into a principal VA model seems nearly impossible given that the vast majority of principals are not observed in multiple schools. As our prior results demonstrate, even separating fixed principal and school effects places considerable constraints on the data. Match effects add yet another layer of complexity. Further, without considerable principal mobility, FE strategies to identify match effects will suffer from a considerable small sample bias (Jackson 2013). Nonetheless, we can examine how the accuracy of principal VA models—which typically assume portability of principal effectiveness across schools—changes when match effects are a large component of the principal effect.

3 Simulation

To examine the accuracy of principal VA from connected network models, we employ a simulation that compares principals' VA estimates to known effects, which are drawn randomly. Simulation approaches have been used previously to examine teacher VA (e.g., Guarino et al. 2015; Guarino, Reckase, and Wooldridge 2015). Different from simulation studies of teacher VA, however, we use administrative data from Tennessee and NYC as the structure of the simulation. This is important because the focus of our study is the connected network approach, and the accuracy of these estimates depends on the network structure of the dataset. Specifically, for both contexts, we start with a dataset at the student-by-year level that contains unique identifiers for student, principal, and school. We then randomly draw the principal and school effects, and generate student outcomes as a function of these effects. Below, we outline our assumed data-generation process and simulation procedures.

3.1 Data-Generating Process

To isolate issues relevant to the connected networks approach, we assume a fairly straightforward data-generating process for student test scores:

$$A_{ijst} = \lambda A_{ijs,t-1} + \theta_{jst} + c_i + e_{ijst} \tag{2}$$

where $A_{ijs,t-1}$ is the prior-year score with a persistence parameter λ, θ_{jst} is the school-by-year-specific contribution to the current-year score, c_i is time-invariant student heterogeneity, and e_{ijst} is a random error term that is assumed to be independent over time. This mirrors the DGP used by Guarino et al. (2015) and Guarino, Reckase, and Wooldridge (2015) in their teacher VA simulations, except that we conceptualize school-level inputs as a single school-by-year effect rather than a set of teacher indicator variables. We refer to the school-by-year effect as "school performance," which is a linear function of fixed principal quality (δ_j), a principal-by-school match effect (α_{js}), fixed school-level factors that the principal cannot control (γ_s), and a school-by-year random shock (v_{jst}):

$$\theta_{jst} = \delta_{j(t,s)} + \alpha_{js(t,s)} + \gamma_s + v_{jst} \tag{3}$$

In this DGP, changes in school performance (aside from yearly random deviations) are completely determined by the principal.

As with any simulation approach, we acknowledge that this DGP is a simplification, and that the true nature of principal effects may be substantially more complex. In particular, we are assuming that principal and school quality are fixed and that unobserved student heterogeneity has a constant effect in each year. We additionally assume no time-varying student or family effects, no interactions between students and principals or schools, and no peer effects. Finally, we assume that test scores have no measurement error and there is no serial correlation in the error term. These simplifications allow us to understand more deeply how the principal and school fixed effects model may or may not produce good estimates of principal quality according to the structure of connected networks. If the models perform poorly here, they likely face greater challenges in the context of real data.

We show our simulation parameters in Table 1. Panel B shows the parameters that are fixed across all simulations, which we chose following Guarino, Reckase, and Wooldridge (2015). Specifically, we assume a persistence parameter of 0.5, implying that past school and family inputs decay geometrically across years (Sass, Semykina, and Harris 2014), though our results are not particularly sensitive to this choice.

Panel C shows the parameters that we vary across simulations: (1) the percentage of variance in student gain scores explained by school performance; (2) the relative magnitude of the principal-by-school match effect; and (3) the correlation between principal quality and the fixed school effect. For the magnitude of the school performance effect, we test scenarios where schools are responsible for 5% or 10% of the total variance in student achievement growth, which corresponds roughly to the range found using variance decomposition methods for math and reading scores in Tennessee and New York City.[3] Across the simulations, we hold constant the relative importance of the principal (45% of the variance of the school performance effect), school (45%), and random shock (10%), but we vary how much of the principal effect is the fixed component (δ_j) versus the match component (α_{js}).[4] Specifically, we test models where there is no match effect, where the match effect is roughly one-quarter of the total principal effect, and where the match effect is slightly larger than the stable principal effect. Finally, we examine different correlations (0.4, 0, and -0.4) between the fixed principal and school effects, where a positive correlation means that effective principals are more likely to work in effective schools.

3.2 Models for Estimating Principal VA

The purpose of our simulation is to compare principals' true effects to their estimated effects from VA models using principal and school fixed effects. Specifically, we estimate:

$$Y_{ijst} = \tilde{\lambda} Y_{ijs,t-1} + \tilde{\delta}_j + \tilde{\gamma}_s + e_{ijst} \tag{4}$$

As previously described, this model produces estimates of principal effects ($\tilde{\delta}_j$) that are relative to the mean of principals within the same connected network. We refer to this model as "principal FE + school FE" (P+S). One challenge is the estimation

[3] These results are shown in Appendix Table C.1. Specifically, we estimate a school-by-year random effects model for current-year test scores with controls for prior-year test scores, student characteristics, and school-by-year means of student characteristics.

[4] Supporting this choice, in the variance decomposition results shown in Appendix Table C.1, we find that the variance components for principals and schools are roughly equal for both math and reading.

of $\tilde{\lambda}$, which is the persistence parameter for prior-year test score, $Y_{ijs,t-1}$. As noted in prior work, this estimate is biased upwards due to the presence of α (fixed student heterogeneity) in the error term (Guarino, Reckase, and Wooldridge 2015; Andrabi et al. 2011). Further, because most students remain in the same school (and have the same principal) between year $t-1$ and t, the upward bias in $\tilde{\lambda}$ leads to attenuation of $\tilde{\delta}$ and $\tilde{\gamma}$. Consistent with prior studies, however, we find that the impact of this bias on the accuracy of VA estimates is small, and we thus proceed with the lagged score model, which is the common approach for estimating principal effects.[5]

In addition to the P+S model, We also examine three alternative specifications. First, to understand the importance of variance inflation introduced by the sparse networks of principals and schools, we estimate models that restrict the size of these networks. Specifically, we replace principal FE with principal-by-school FE, which we refer to as "principal-school + school FE" (P-S+S). By estimating an effect for each principal-by-school spell rather than each principal, connected networks are restricted to a single school—principals cannot be compared across schools. While this greatly limits the comparison set for many principals, it also reduces the noise component inherent to the principal and school FE model.

Second, we estimate a model with principal-by-district FE and school FE (P-D+S). P-D+S is a middle ground between P+S and P-S+S, allowing networks to be as large as single district instead of limiting them to a single school. This model addresses a concern that two disconnected districts may be linked by just a single principal who moves between them, essentially creating a single connected network of the two districts. Such an occurrence can create an extreme bottleneck that drives variance inflation in the VA estimates. In essence, P-D+S is a method of attempting to eliminate edges in connected networks that are most likely to represent bottlenecks.

[5] In the case of teacher VA, models typically do not include school FE and students have new teachers each year. Thus, the teacher effect is not a structural component of $Y_{ijs,t-1}$. In the case of principals, both the school and principal contribute to the prior-year score. Any bias in $\tilde{\lambda}$, then, will necessarily lead to bias in principal and school effects. However, because principals and schools have *continued* impacts on student achievement and the prior-score is a good proxy for unobserved student heterogeneity, the bias does not substantially diminish the accuracy of the principal VA estimates.

Finally, we estimate a model that does not include school FE, which we call "principal FE only" (PO). This is effectively school value-added averaged over the principal's tenure in the school. While we anticipate that this model will perform poorly in scenarios where the fixed school component is a large contributor to school performance, omitting school FE avoids the estimation challenges endemic to the connected networks approach and allows for a global ranking of principals. In particular, this models helps us to understand whether the bias/precision tradeoff makes sense when choosing to include school fixed effects in principal VA models.

3.3 Assessing Model Performance

For each of the 18 unique combinations of the simulation parameters in Table 1, we run 10 Monte Carlo replications of the simulation, where the principal, match, school, and student effects are drawn randomly. Given the large-scale nature of our datasets, our results are highly consistent across replications. In the performance metrics described below, we report the simple average across the 10 replications.

We consider two main questions about the performance of principal VA models. (1) How accurately do VAMs rank principals according to their true effects? (2) To what extent does the magnitude of principal VA accurately reflect the true magnitude of principals' effects? To answer these questions, we draw on simulated data to compare the VA estimates to the true principal effects. In the models that contain school FE, however, we must first adjust the true principal effects by centering them within connected networks.[6] Importantly, we define the true principal effect to include both the fixed and match components of the principal effect.[7] While there are certain scenarios where isolating the fixed component of principal quality is desirable (e.g., wanting

[6] Specifically, we residualize true principal effects on the vector of network FE. For the P-S+S and P-D+S approaches, we effectively treat each principal-by-school or principal-by-district spell as a separate principal, though our results are essentially identical if we weight our performance metrics inversely by the number of spells per principal.

[7] For principals who work in multiple schools, we construct a time-invariant total principal effect that uses a weighted average (according to the number of student-by-year observations) of the match components from each of their schools.

to understand whether a principal is likely to be effective in a different school), our primary aim is to evaluate models that attempt to measure the principal's overall contribution to improving student achievement in the schools where they actually worked, regardless of whether it reflects a fixed or match effect.

We then report four summary measures. First, we compute the Pearson correlation between the principal VA estimates and the true principal effects.[8] A correlation of one, to be specific, would indicate that the model perfectly ranks principals in terms of their true effectiveness (within networks). Low correlations between estimated and true principal effects may be a product of bias, imprecision, or both. Thus, we also report performance measures that isolate these factors. Our second measure captures bias in the VA estimates by estimating a simple regression of VA estimates as a function of true effectiveness:

$$\tilde{\delta_j} = \beta \delta_j + e_j \tag{5}$$

where $\tilde{\delta_j}$ is principal j's VA estimate and δ_j is their true effect. $\beta = 1$ would indicate that the VA model produces unbiased estimates of principals' true effects. If $\beta > 1$ ($\beta < 1$), the VA estimates amplify (condense) the true principal effects. A model that produces VA estimates where $\beta > 1$ may correctly rank principals even though the estimates are systematically biased. Note that because $\tilde{\delta_j}$ is on the left-hand side of the equation, estimation error will not lead to an attenuation of β.

Third, to understand the degree of variance inflation in the VA estimates, we report the ratio of the standard deviations of the estimated and true principal effects: $\sigma_{\tilde{\delta_j}}/\sigma_{\delta_J}$. A ratio of one would indicate that the distribution of principal VA estimates provides a good approximation of the magnitude of principal effects on student outcomes, while a ratio greater than (less than) one would indicate that the model overstates (understates) the magnitude of principal effects. In addition to capturing variance inflation from estimation error, however, this ratio also captures downward bias in the variance

[8] Using Spearman rank correlations yields very similar estimates.

due to small networks. Thus, to isolate the estimation error, our fourth summary measure is the ratio of the standard deviation (SD) of the estimated principal effects and the true principal effects after demeaning them within the connected networks.

3.4 Data and Network Structure

We apply the DGP described by equation 2 to actual student-level administrative data from Tennessee and NYC. For both datasets, we can identify the linkages between students, schools, and principals, which allows us to test VA models using the actual connected networks of principals. For Tennessee, the analysis years run from 2007–2019 and include 3,835 principals, 1,719 schools, and roughly 5.1 million student-by-year observations. The NYC data goes from 2003–2017, containing 3,144 principals, 1,323 schools, and roughly 6.0 million student-by-year observations.

Table 2 summarizes the connected networks of principals in Tennessee and NYC, respectively. In Tennessee, there are 762 individual networks, though most either consist of a single school (74%) or are small (22.8%) which we define as a network with between two and five schools. Single-school networks have 2.5 principals while small networks have, 5.8 principals and 2.6 schools, on average. Tennessee also has 22 medium-sized networks (6–15 schools), comprising less than 3% of all individual networks, and two large networks (16+ schools). Six percent of principals have no network, meaning that they were the sole principal observed in a school across the analysis years.

By contrast, NYC has fewer principals in medium- or large-sized networks despite our access to a longer panel, and 60% of principals are in a single-school network. The average number of principals and schools in the single and small-sized networks in NYC are very similar to those found in Tennessee, as are the proportions of principals who have no network. NYC's lack of principals in medium- and large-sized networks reflects an important difference in the principal labor markets between Tennessee and NYC; while we do observe some principals who transfer schools in Tennessee (which creates the necessary linkages for larger networks), the principal transfer rate in NYC is nearly zero. In other words, principals who leave their positions in NYC almost exclusively transfer out of the district (movement which we cannot observe with these data) or move out of the principalship entirely (e.g., retirements).

The final three rows of Table 2 concern the connectivity structure of the networks, which determines the precision of the FE estimates from the P+S model. Specifically, principals' VA estimates become more reliable as the network grows more dense, both in terms of the number of direct comparisons between different principals as well as the number of observations (i.e., student test scores). Following the approach of Jochmans and Weidner (2019), we analyze each network as a bipartite graph to produce the predicted amount of variance inflation (reported as a percentage of the error variance) based on its normalized Laplacian matrix. We also report the smallest nonzero eigenvalue (λ_2)—a global measure of connectivity where $\lambda_2 \to 0$ as the network becomes more sparse. Across all network sizes in both contexts, principal VA models benefit from large sample sizes, since all of the tested students in a school will contribute to estimating the principal's effect. Larger networks tend to be weakly connected, however, which will lead to non-trivial variance inflation in the VA estimates. In Tennessee's two large networks, for instance, the mean predicted variance inflation is 0.027, which is scaled in terms of the error variance of student test score growth.

To make more concrete the concept of network connectivity, Figure 1 shows two examples of medium-sized principal networks in Tennessee. In each plot, the numbered nodes represent principals, with edges representing comparisons among principals who worked in the same school. As an example, the left part of plot A shows that principals 2, 3, and 11 worked in the same school and can be directly compared. Principal 3 also worked in a (different) school with principals 7 and 16. This allows for indirect comparisons between principals in these subsets. As more principals switch schools (as opposed to leaving the principalship altogether), networks will grow larger and more indirect comparisons will become possible.

While the networks in plots A and B are similar in the number of principals and schools, their connectivity (summarized by λ_2, the smallest non-zero eigenvalue) differs.

Intuitively, a network is weakly connected when it is easy to separate it into two substantial sub-networks by removing edges. This is the case for plot A, where there are fairly few edges linking the two sides, and these centralized edges have relatively low weights (denoted in the plot by the edge width). In plot B, there are far more edges, which represents greater mobility among schools. There are also redundancies such that the network cannot be split into two sub-networks by removing a single principal. As a result, variance inflation in the network shown in plot A is predicted to be roughly four times greater than in the network shown in plot B.

Overall, Tables 2 shows that the precision of VA estimates is very high for principals in small or medium-sized networks in Tennessee and NYC. Of course, the trade-off is that these VA estimates are highly localized measures of performance—the majority of principals can only be compared to other principals who worked in the same school. While there are a few large networks that, in theory, produce a more globalized measure of performance, the underlying network structure is extremely weak, leading to VA estimates that contain considerable estimation error.

4 Results

Figures 2 and 3 show results for the four summary measures across simulations in Tennessee and NYC, respectively. For the sake of parsimony, we only report the simulation results where school performance accounts for 5% of the total variance in student test score growth, with the full results available in Appendix B. In each panel, the performance measure (averaged across the 10 iterations) for each model specification (defined by the legend at the bottom) is shown by the size of the principal–school match (top x-axis) and the correlation between the fixed principal and school effects (bottom x-axis).

Panel A shows that the accuracy of principal VA estimates varies substantially according to both model specification and the magnitude of the principal–school match effect. We begin with the "no match" scenario—the ideal scenario for the connected networks approach where systematic deviations in school performance are completely determined by the fixed principal effect. Here, the P+S, P-D+S, and P-S+S models all perform relatively well, producing correlations with true principal effects that are above 0.7 in Tennessee and just under 0.9 in NYC. P-S+S, however, is always the most accurate, though it faces the limitation that VA estimates are highly localized—principals can only be compared to others who worked in the same school. In NYC, the difference between the P+S and P-S+S models is smaller. Because the vast majority principals are already in small networks, limiting to single-school networks has less of an impact on the results relative to Tennessee.

Given that the inclusion of school FE in principal VA models creates challenges—either in the form of noisy estimates or limited comparison sets—a clear question is whether to omit school FE entirely. The results in Figures 2 and 3 demonstrate, however, that if there is a fixed school contribution to test score growth, the exclusion of school FE (PO model) can lead to wildly inaccurate VA estimates. This model is particularly inaccurate when the true principal and school effects are negatively correlated.

The sensitivity of the estimates to the inclusion of school FE is due to the fact that most principals are observed in only a single school. Any omitted school effect will be incorrectly attributed to the principal, producing an inaccurate estimate even if there is no absolute sorting of principals to schools (i.e., where the correlation between the fixed principal and school effect is zero). In essence, the issue is one of omitted variables bias. The connected networks model is, effectively, a model with two large sets of indicator variables that are highly correlated with one another. Omitting the school indicators creates bias in each of the principal indicators, which grows in size with the magnitude of school effects.

Concerning our second research question, Panel B of Figure 2 shows that the SD of the distribution of principal VA estimates often does not correspond to the true SD of principal effects. As with panel A, the results vary both by model and the size of the principal–school match. For the P+S approach, the SD of principal VA consistently overstates the magnitude of principal effects, with greater bias when the principal–school match effect is larger. By contrast, the P-S+S model *understates* the importance of principals. As we outlined earlier, there are two sources of bias in the estimated magnitude. First, VA estimates from large networks contain substantial estimation error, leading to variance inflation that is evident in the P+S results. Second, VA estimates from small networks understate principal effects because the school fixed effect captures variation in true principal quality. In New York City, the dynamic is again slightly different because of the large number of principals in small networks. Here, both P+S and P-S+S understate the magnitude of principal effects, as the downward bias from small networks outweighs the upward bias from estimation error in large networks.

Panels C and D of Figures 2 and 3 help demonstrate why the P+S and P-D+S models are less accurate than the P-S+S model. Panel C shows the results of regressing the principal VA estimates on the true principal effects. In models with school FE, the estimated coefficient is roughly 0.9, with a small amount of bias introduced by controlling for the prior-year test score in models with principal and school effects. Panel D, however, shows that the P+S model contains substantially greater estimation error. This noise component contains no information about principal effectiveness and thus lowers the accuracy of the P+S model. As expected, the P-D+S model in Tennessee is bounded between the P+S and P-S+S models, as it partially restricts network sizes, though not to the extent of the P-S+S model.

We next consider the accuracy of principal VA in the presence of principal–school complementarities. Figures 2 and 3 show that even a small match effect undermines the correlation with the true principal effects. By contrast, match effects have negligible (PO) or no (P-S+S) influence on the other models. This inaccuracy is a product of additional noise in the principal effect estimates rather than a systematic bias. Because the P+S model relies on mobile principals to identify the school effects, the addition of a principal-school match effect makes this source of variation inherently unreliable. Put another way, using the performance of principal A in two different schools to make indirect comparisons among other principals is problematic if principal A's true performance varies in these two schools, particularly if principal A is the *only* connection between the schools. In short, principal–school complementarities amplify the existing weakness of the connected networks approach for estimating principal effects.

The P-S+S and PO models, which do not rely on this form of mobility for identification, are relatively unaffected by the existence of match effects. The P-S+S model, by definition, includes any match effects because it produces a separate estimate for each principal-school combination. The PO model will simply provide an estimate of the weighted average of principal effectiveness (i.e., the portable component and match effect) across the schools in which a principal worked.

To reinforce the importance of network structure, we next compute our performance measures for the P+S model according to network size: single (1 school), small (2–5 schools), medium (6–15 schools), and large (16+ schools). Given the similar patterns for Tennessee and NYC and the limited variation in network size in NYC, we focus our discussion on the Tennessee results.[9] Consistent with the network structure analysis in Table 2, Figure 4 shows that the estimation error in the P+S model is driven by larger networks. Specifically, Panel A shows that the correlation between estimated and true principal effects from large networks are much lower than those from smaller networks. In the "no match" scenario, correlations from small networks are roughly 0.9, while those from large networks are roughly 0.6.

Panels C and D show that this difference in accuracy between larger and smaller networks is completely driven by variance inflation. While there are no substantive differences in the bias measure in panel C, the ratio of standard deviations in panel D are substantially greater for large networks. In the "no match" scenario in panel D, for example, the estimated SD of principal VA for single or small networks is approximately equal to the true within-network SD of principal effects, while the estimated SD for large networks is roughly 1.6 times larger than the true SD.

[9] The NYC results are shown in Appendix Figure A.1.

The results in Figure 4 demonstrate a key tradeoff in the connected networks approach. As the size of the connected networks grows, principal VA estimates become less localized, which increases their usefulness as a measure of principal effectiveness. At the same time, the reliability of the VA estimates decreases due to weak network structure. Thus, the ability to link an increasing number of principals is not unambiguously beneficial even if the underlying assumptions of the principal and school fixed effects model are met.

4.1 Shrinkage Estimators

The previous section demonstrates that FE estimates of principal quality from connected networks models contain substantial estimation error in large networks, which leads to low correlations with the true principal effects and inflated estimates of the magnitude of principal effects. These inaccuracies are further magnified in the presence of principal–school complementarities.

One way to potentially reduce the variability of principal VA estimates is to implement a shrinkage estimator, such as Empirical Bayes's (EB), that adjusts for the estimation error in the principal FE. The intuition of the EB approach, in this case, is that principals with very high (low) VA estimates are likely suffering from positive (negative) estimation error. EB estimation accordingly shrinks these estimates toward the mean principal effect, yielding a biased but less noisy VA estimate. In theory, the shrunken estimates should have a higher correlation with the true principal effects.

We implement the EB approach in two ways. The first is to make a post hoc adjustment to the estimated FE by drawing on their standard errors as a measure estimation error, according to the following formula:

$$\hat{\delta}_j^{EB} = \lambda_j \hat{\delta}_j^{FE} + (1 - \lambda_j)\bar{\delta} \tag{6}$$

The FE estimate $\hat{\delta}_j^{FE}$ is shrunken towards the mean principal effect ($\bar{\delta} = 0$) by the factor $1 - \lambda_j = 1 - \frac{\hat{\sigma}_\delta^2}{\hat{\sigma}_\delta^2 + \hat{\zeta}_j}$. This shrinkage factor is a function of the estimated variance of principal effects ($\hat{\sigma}_{\delta}^2$) and the estimated error variance of principal j's effect. The latter quantity is simply the squared standard error of the FE estimate for principal j, while the former is approximated by the mean of the square of the standard errors of $\hat{\delta}_j$ subtracted from the variance of $\hat{\delta}_j$ (e.g., Aaronson, Barrow, and Sander 2007; Branch, Hanushek, and Rivkin 2012).[10] Intuitively, as error in a principal's FE estimate ($\hat{\zeta}_j$) increases relative to the variance of principal effects ($\hat{\sigma}_{\delta}^2$), the shrinkage factor $(1 - \lambda_j)$ increases, pulling the EB estimates towards zero.

Our second approach to implement EB is to estimate a mixed model where principal and school are random effects instead of fixed effects. From this model we obtain the best linear unbiased predictions (BLUPs) for the principal effects. To ease computational demands, we limit our shrinkage exercise to the largest connected network in Tennessee, which contains 568 principals and 228 schools. This network also has the largest amount of variance inflation, making it a useful test case for the EB approach. We focus on the DGP where the correlation between the fixed principal and school effects are zero, conducting 50 iterations across each of parameters for the magnitude of the principal–school match effect, using the same performance measures as the baseline simulation.

Figure 5 shows the simulation results for the FE and two EB approaches. Panel A shows that the mixed model substantially improves upon the estimates from the P+S model, while the shrunken FE approach does not. For example, the P+S estimates in the "no match" scenario are correlated with the true principal effects at only 0.52, which increases to 0.54 using shrunken FE but to 0.85 for the BLUPs from the mixed model. In the P-S+S and PO models, there is virtually no difference between the FE estimates and either of the shrunken estimates. As shown in panel B, the mixed model also produces a remarkably accurate estimate of the magnitude of principal effects, even in the presence of a large principal–school match component. By contrast, the shrunken FE still overstates the SD of principal effects, particularly when match effects

[10] We obtain the standard errors for the FE using the routine proposed by Mihaly et al. (2010), which accounts for the sum-to-zero constraints within connected networks.

exist.

The differences between the two shrinkage approaches stem from the fact that, in the shrunken FE model, the school effects are included as covariates and are estimated via the FE estimator versus the random effects (RE) estimator. While FE produces consistent estimates of the school effects, these estimates are plagued by the same estimation error as the principal effects. In essence, the shrunken FE approach does little to address the root cause of the estimation error, which is the collinearity of principal and school assignment. The RE model, which assumes (incorrectly) that principal and school assignments are uncorrelated, produces biased, yet substantially more precise, estimates of the principal and school effects. The net result is that the bias/variance tradeoff is squarely in favor of the RE approach when the source of estimation error is the collinearity of the principal and school effects (i.e., the P+S model). In the P-S+S case, however, there is little estimation error because of the constrained network sizes. Here (and with PO), the two shrinkage approaches are more or less identical. With the large number of students contributing to each principal's estimated effect, the FE and shrinkage estimates are very similar.

Panels C and D demonstrate the bias/variance tradeoff more clearly. In the P+S model, both shrinkage procedures reduce estimation error, which yields a SD of principal VA close to the true SD of principal effects. This shrinkage, however, results in biased estimates, whereby differences in shrunken VA understate the true differences in principal effectiveness. In the case of shrunken FE, the gain in precision is almost perfectly offset by the bias, while the bias in the mixed model is much smaller. In the P-S+S and PO models, the FE estimates are already very precise, such that there is little to be gained through shrinkage.

4.2 Empirical Application

As a final supplement to our simulation results, we provide a brief demonstration using principal VA estimates from actual student test scores. Specifically, we estimate the P+S, P-S+S, and mixed models, with the distribution of VA estimates summarized in

Table 3.[11] In addition to the full sample, we show the SD of the estimates for principals by network size. We again focus on the Tennessee results, with the New York City results shown in Appendix Table A.1.

As our simulation results would suggest, the SD of the P+S model estimates for actual math and reading VA increases substantially with network size, reflecting the variance inflation due to sparse network structure. For instance, the SD of math VA in single-school networks is 0.09, meaning that a 1 SD increase in principal VA raises math test scores by 0.09 student-level SD, on average. In large networks, this SD increases to 0.32. Turning to the mixed and P-S+S models, which should contain little to no estimation error, we observe that the estimated magnitude of principal VA does increase in larger networks, but by substantially less than would be implied by the P+S results. The mixed model—which in our simulations provides the most accurate estimates of the magnitude of principal VA—suggests that the true SD in large networks is 0.14 for math and 0.06 for reading. As previously noted in our simulation results, the within-network SD of principal VA in small networks will understate the true SD of principal quality, as the school FE sweeps out real differences in average principal quality between schools.

In an effort to examine the accuracy of principal VA across these specifications, we compare VA to alternative measures of principal performance. In Tennessee, principals beginning in 2011–12 receive rubric-based ratings from their supervisors as part of the state's high-stakes educator accountability system, where the average score comprises 50% of a principal's summative evaluation rating. Prior work has documented positive, though weak, relationships between supervisor ratings and principal VA (Bartanen 2020; Grissom, Blissett, and Mitani 2018). We hypothesize, however, that these correlations may be somewhat attenuated due to estimation error in the principal VA

[11] In the control vector, we include a broader set of student and school characteristics to align with common principal VA specifications. Specifically, we control for cubics of prior-year test scores in math and reading, prior-year attendance rate, student demographic characteristics (race/ethnicity, gender, economically disadvantaged, English learner, gifted classification, special education classification, and grade repetition), and school-by-year averages of these demographics.

estimates.[12]

Figure 6 shows binned scatterplots for predicting a principal's mean supervisor rating as a function of their math or reading VA estimate across model specifications. We additionally control for the relevant network fixed effects to maintain the within-network interpretation of the VA estimates. The left plots show how large estimation error attenuates the correlation between P+S VA estimates and supervisor ratings. While there is an apparent upward slope when restricting to principals in the middle of the distribution, the high-leverage observations at the tails (i.e., principals with very high or low VA estimates) flatten the estimated regression line. That principals at the extremes of the VA distribution have supervisor ratings close to the mean further suggests that their VA estimate is suffering from positive or negative estimation error.

Estimation error can be reduced through a shrinkage approach (i.e., the mixed model) or by limiting the size of connected networks (i.e., the P-S+S model). In both approaches, the distribution of VA is compressed and the estimated correlations with supervisor ratings increase. This provides empirical support for the simulation results, which show that estimates from the mixed model or P-S+S model will produce more accurate rankings of principals' true effects. Finally, Appendix Figure A.2 shows results from an equivalent exercise that predicts residualized student test scores in one subject as a function of principal VA in the other subject (e.g., predicting math scores using a principal's reading VA).

5 Conclusion

There is a growing interest in using value-added models to estimate principals' contributions to student outcomes. Accurately isolating principal effectiveness requires accounting for school-level factors that affect student achievement but are not controlled by principals. The common approach to address this issue is to estimate a

[12] While principals receive an average observation score each year, we construct a time-invariant measure that averages across all available years. When comparing to the principal-by-school VA estimates, we limit to observation scores that were received in that same school.

"connected networks" model with principal and school fixed effects. The accuracy of this model, however, has not been rigorously tested. Specifically, the inclusion of school FE—while conceptually important for mitigating bias—creates challenges with respect to the reliability of principal VA estimates.

Using simulated test scores applied to the actual principal–school assignments across long panels from Tennessee and New York City, we reach several important findings. First, limited mobility of principals combined with high rates of attrition makes the connected networks model difficult to implement. There is insufficient variation to jointly identify both principal and school effects, such that principal VA estimates are either highly localized—reflecting performance relative to only a handful of other principals—or very imprecise. In both Tennessee and New York City, the modal principal is in a single-school connected network. While estimates from small networks are reliable, they are less useful from a practical perspective and they understate the magnitude of principals' effects. On the other hand, VA estimates from large networks reflect a principal's performance relative to a much larger group, but the underlying network structure is weak. As a result, VA estimates from large networks are unreliable and their variance overstates the magnitude of principals' effects. The precision of VA estimates in large connected networks can be improved by employing a shrinkage estimator, though our simulation results suggest that a mixed model performs substantially better than a post-hoc shrinkage of the principal fixed effects.

Our results help to inform the estimation of principal VA. Those implementing principal VA models should consider the intended use of the estimates when choosing the most appropriate specification. For most applications, models with principal random effects and school FE are the best available option. This is particularly true when the magnitude of principal effects is the main parameter of interest. While, models with principal-by-school and school FE produce VA estimates with stronger internal validity and higher reliability, they understate the importance of principal quality and the measures are highly localized. The inability to compare principals across schools, in particular, makes this model unfavorable for accountability purposes. A final alternative is to omit school FE entirely from the model. While this alleviates variance inflation and small network issues, it comes with a risk of substantial bias in the VA estimates, as any unobserved school-level factors that are not completely captured by observable school characteristics will be mistakenly attributed to principals.

Given our results, a clear question is whether principal VA can really provide useful information about principal effectiveness. Certainly, the multi-faceted and indirect nature of principals' contributions to student outcomes makes estimating principal effects a formidable challenge. We stress, however, that even imperfect principal VA models may contain valuable information that is not captured by alternative measures. A clear strength of the principal effects literature, for instance, is the ability to avoid penalizing principals who work in the most challenging schools. This is particularly important given evidence that rubric-based ratings of principal practice and school value-added—two commonly used alternative measures of principal performance—in part hold principals accountable for factors they cannot control (Chiang, Lipscomb, and Gill 2016; Grissom, Kalogrides, and Loeb 2015; Grissom, Blissett, and Mitani 2018). As a final caution, we note that our simulation analysis makes a number of assumptions about the nature of principal effects on student outcomes. These assumptions—most notably, that new principals can immediately change school performance and their effects are fixed over time—are helpful for isolating issues related to the connected networks approach, but may not hold in practice. While this study is an additional step in understanding the extent to which principal VA models can provide accurate estimates of principals' effects, there is a continued need for work that rigorously examines their validity and reliability.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and student achievement in the Chicago public high schools." *Journal of Labor Economics* 25 (1): 95–135.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc. 2011. "Do value-added estimates add value? Accounting for learning dynamics." *American Economic Journal: Applied Economics* 3 (3): 29–54.

Bartanen, Brendan. 2020. "Principal Quality and Student Attendance." *Educational Researcher* 49 (2): 101–113.

Bartanen, Brendan, and Jason A Grissom. 2021. "School Principal Race, Teacher Racial Diversity, and Student Achievement." *Journal of Human Resources.*

Branch, Gregory F., Eric A. Hanushek, and Steven G. Rivkin. 2012. "Estimating the Effect of Leaders on Public Sector Productivity: The Case of School Principals." Cambridge, MA.

Burkhauser, Susan. 2017. "How Much Do School Principals Matter When It Comes to Teacher Working Conditions?" *Educational Evaluation and Policy Analysis* 39 (1): 126–145.

Chiang, Hanley, Stephen Lipscomb, and Brian Gill. 2016. "Is School Value Added Indicative of Principal Quality?" *Education Finance and Policy* 11 (3): 283–309.

Coelli, Michael, and David A Green. 2012. "Leadership effects: school principals and student outcomes." *Economics of Education Review* 31:92–109.

Dhuey, Elizabeth, and Justin Smith. 2018. "How school principals influence student learning." *Empirical Economics* 54:851–882.

Grissom, Jason A., Demetra Kalogrides, and Susanna Loeb. 2015. "Using Student Test Scores to Measure Principal Performance." *Educational Evaluation and Policy Analysis* 37 (1): 3–28.

Grissom, Jason A, and Brendan Bartanen. 2019. "Strategic Retention: Principal Effectiveness and Teacher Turnover in Multiple-Measure Teacher Evaluation Systems." *American Educational Research Journal* 56 (2): 514–555.

Grissom, Jason A, Brendan Bartanen, and Hajime Mitani. 2019. "Principal Sorting and the Distribution of Principal Quality." *AERA Open* 5 (2): 1–21.

Grissom, Jason A, Richard S. L. Blissett, and Hajime Mitani. 2018. "Evaluating School Principals: Supervisor Ratings of Principal Practice and Principal Job Performance." *Educational Evaluation and Policy Analysis* 40 (3): 446–472.

Guarino, Cassandra M., Michelle Maxfield, Mark D. Reckase, Paul N. Thompson, and Jeffrey M. Wooldridge. 2015. "An Evaluation of Empirical Bayes's Estimation of Value-Added Teacher Performance Measures." *Journal of Educational and Behavioral Statistics* 40 (2): 190–222.

Guarino, Cassandra M, Mark D Reckase, and Jeffrey M Wooldridge. 2015. "Can Value-Added Measures of Teacher Performance Be Trusted?" *Education Finance and Policy* 10 (1): 117–156.

Hallinger, Philip, and Ronald H Heck. 1998. "Exploring the Principal's Contribution to School Effectiveness: 1980-1995." *School Effectiveness and School Improvement* 9 (2): 157–191.

Husain, Aliza N., David A. Matsa, and Amalia R. Miller. 2018. "Do Male Workers Prefer Male Leaders? An Analysis of Principals' Effects on Teacher Retention." *NBER Working Paper Series:* 38.

Jackson, C Kirabo. 2013. "Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence from Teachers." *The Review of Economics and Statistics* 95 (4): 1096–1116.

Jacob, Brian. 2011. "Do Principals Fire the Worst Teachers?" *Educational Evaluation and Policy Analysis* 33 (January): 403–434.

Jochmans, Koen, and Martin Weidner. 2019. "Fixed-Effect Regressions on Network Data." *Econometrica* 87 (5): 1543–1560.

Kline, Patrick, Raffaele Saggio, and Mikkel Sølvsten. 2018. "Leave-out estimation of variance components."

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff. 2015. "Value-added modeling: A review." *Economics of Education Review* 47:180–195.

Kraft, Matthew A., William H. Marinell, and Darrick Yee. 2016. "School Organizational Contexts, Teacher Turnover, and Student Achievement: Evidence from Panel Data." *American Educational Research Journal* 53 (5): 1411–1449.

Kristof-Brown, Amy L, Ryan D Zimmerman, and Erin C Johnson. 2005. "Consequences of Individuals' Fit At Work: A Meta-Analysis of Person-Job, Person-Organization, Person-Group, and Person-Supervisor Fit." *Personnel Psychology* 58:281–342.

Mansfield, Richard K. 2015. "Teacher Quality and Student Inequality." *Journal of Labor Economics* 33 (3): 751–788.

Mihaly, Kata, Daniel F. McCaffrey, J.R. Lockwood, and Tim R. Sass. 2010. "Centering and reference groups for estimates of fixed effects: Modifications to felsdvreg." *The Stata Journal* 10 (1): 82–103.

Mihaly, Kata, Daniel McCaffrey, Tim R. Sass, and J. R. Lockwood. 2013. "Where you come from or where you go? Distinguishing between school quality and the effectiveness of teacher preparation program graduates." *Education Finance and Policy* 8 (4): 459–493.

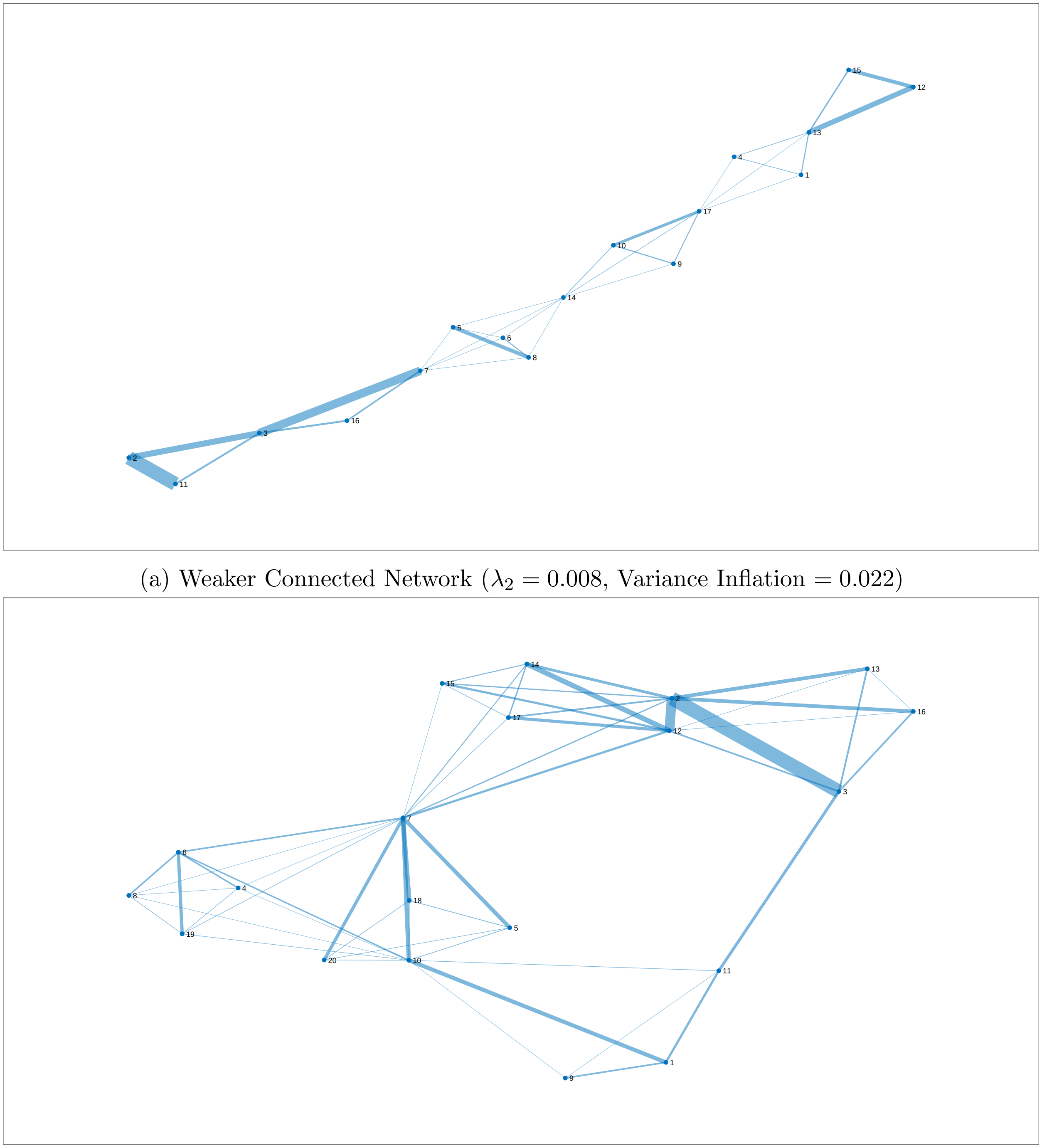
Miller, Jason M, Peter Youngs, Frank Perrone, and Erin Grogan. 2020. "Using Measures of Fit to Predict Beginning Teacher Retention." *The Elementary School Journal* 120 (3): 399–421.

Player, Daniel, Peter Youngs, Frank Perrone, and Erin Grogan. 2017. "How principal leadership and person-job fit are associated with teacher mobility and attrition." *Teaching and Teacher Education* 67:330–339.

Sass, Tim R, Anastasia Semykina, and Douglas N Harris. 2014. "Value-added models and the measurement of teacher productivity." *Economics of Education Review* 38:9–23.

Sebastian, James, and Elaine Allensworth. 2012. "The Influence of Principal Leadership on Classroom Instruction and Student Learning: A Study of Mediated Pathways to Learning." *Educational Administration Quarterly* 48 (4): 626–663.

Verdier, Valentin. 2018. "Estimation and Inference for Linear Models with Two-Way Fixed Effects and Sparsely Matched Data." *The Review of Economics and Statistics:* 1–38.

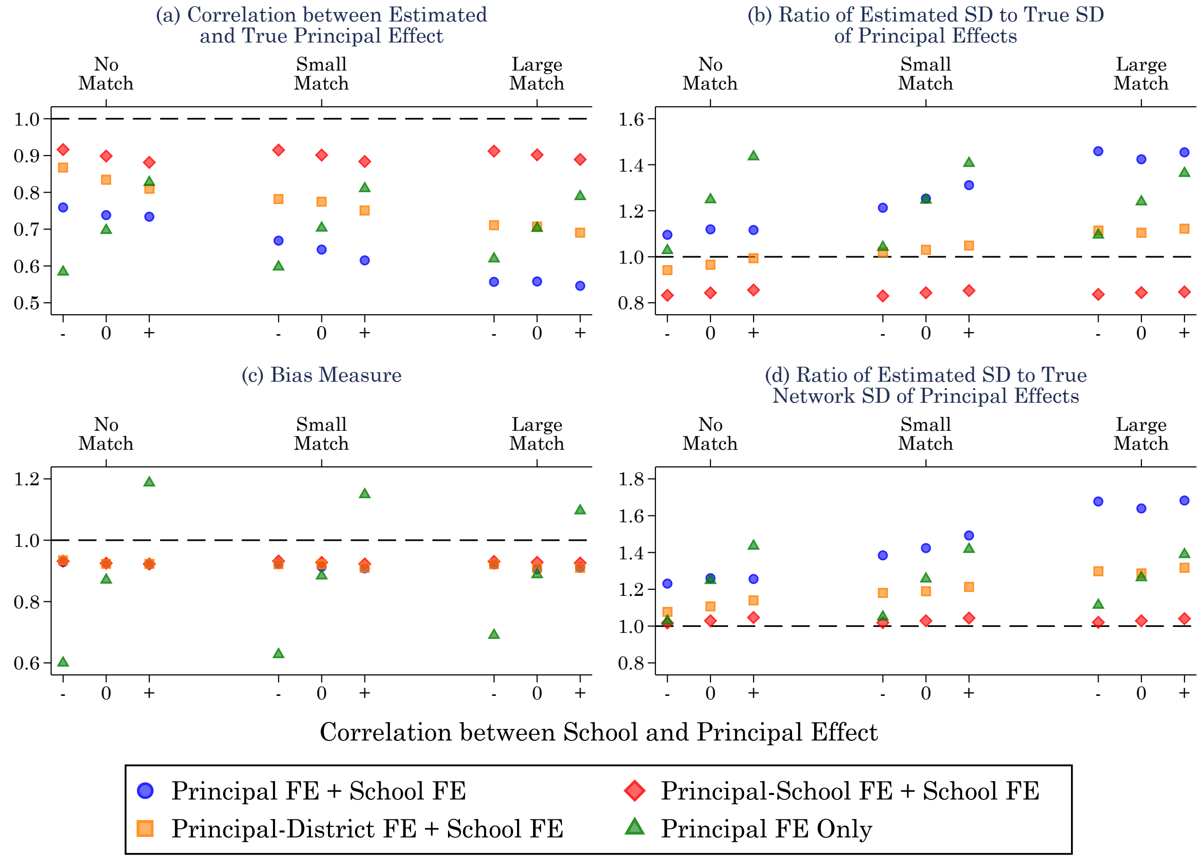


(a) Weaker Connected Network ($\lambda_2 = 0.008$, Variance Inflation = 0.022)

(b) Stronger Connected Network ($\lambda_2 = 0.114$, Variance Inflation = 0.005)

Figure 1: Examples of Connected Networks of Principals

Notes: Each plot shows a single connected network of principals from Tennessee. Nodes represent principals and edges are formed by principals who worked in the same school. In panel A, there are 20 principals across 6 schools. In panel B, there are 17 principals across 6 schools. The weight (shown visually by width) of the edge is determined by the harmonic mean of the number of students that contribute to estimating each principal's effect in the relevant school. Both connectivity (λ_2) and variance inflation are calculated following the approach of Jochmans and Weidner (2019). λ_2 is the smallest non-zero eigenvalue from the normalized Laplacian matrix that corresponds to each connected graph of principals. A smaller eigenvalue indicates that principals in a network are more weakly connected. Variance inflation is expressed in terms of the error variance of student test score growth.



(a) Correlation between Estimated and True Principal Effect

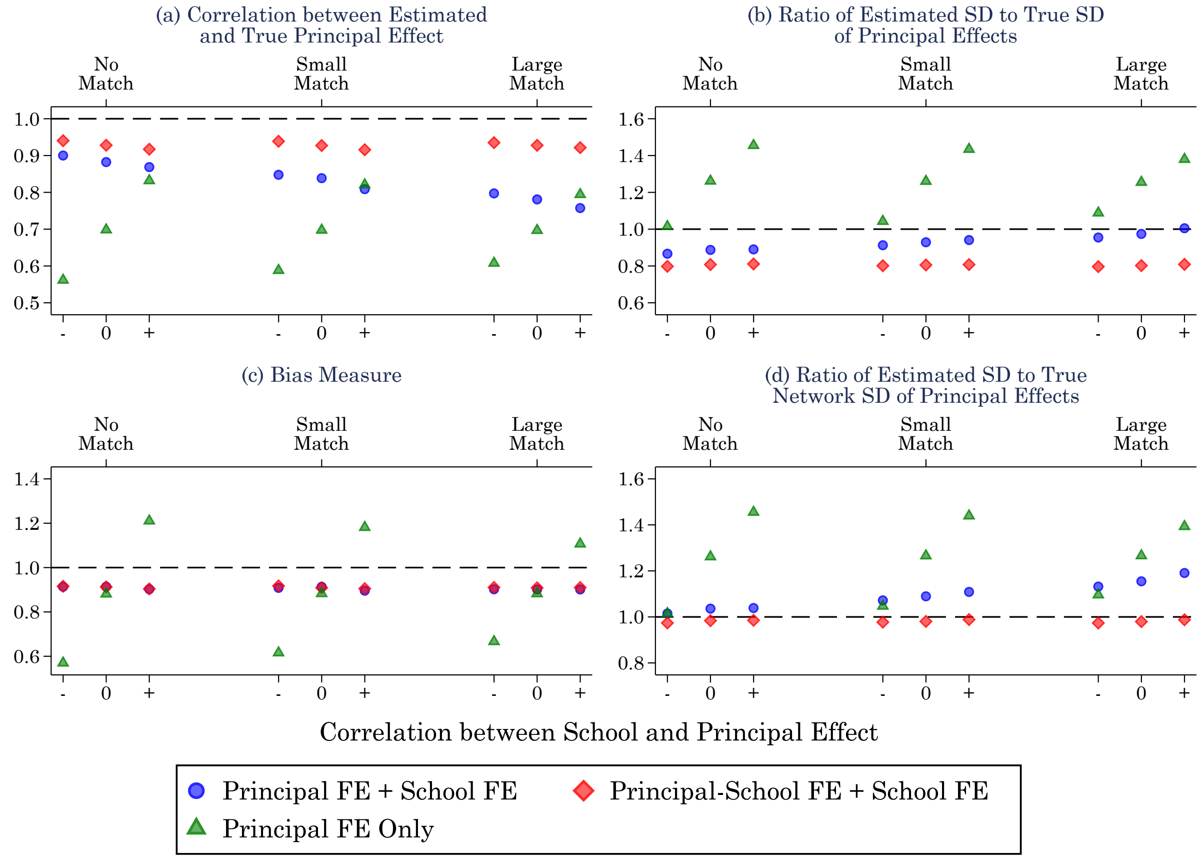
(b) Ratio of Estimated SD to True SD of Principal Effects

(c) Bias Measure

(d) Ratio of Estimated SD to True Network SD of Principal Effects

Figure 2: Baseline Simulation Results (Tennessee)

Notes: In each plot, the y-axis is defined by the header. The bottom x-axis corresponds to the correlation between principal quality and the fixed school effect (-0.4, 0, 0.4). The top y-axis corresponds to the magnitude of the principal–school match effect (0%, 22%, 56% of the total principal effect). In the results shown here, the school performance effect constitutes 5% of the total variance in student test score growth. The full set of simulation results are shown in Appendix B. In Panels A, B, and C, the true principal effects are residualized on network fixed effects corresponding to the connected networks formed by principals and schools for the given VA measure.



(a) Correlation between Estimated and True Principal Effect

(b) Ratio of Estimated SD to True SD of Principal Effects

(c) Bias Measure

(d) Ratio of Estimated SD to True Network SD of Principal Effects

Figure 3: Baseline Simulation Results (New York City)

Notes: In each plot, the y-axis is defined by the header. The bottom x-axis corresponds to the correlation between principal quality and the fixed school effect (-0.4, 0, 0.4). The top y-axis corresponds to the magnitude of the principal–school match effect (0%, 22%, 56% of the total principal effect). In the results shown here, the school performance effect constitutes 5% of the total variance in student test score growth. The full set of simulation results are shown in Appendix B. In Panels A, B, and C, the true principal effects are residualized on network fixed effects corresponding to the connected networks formed by principals and schools for the given VA measure.

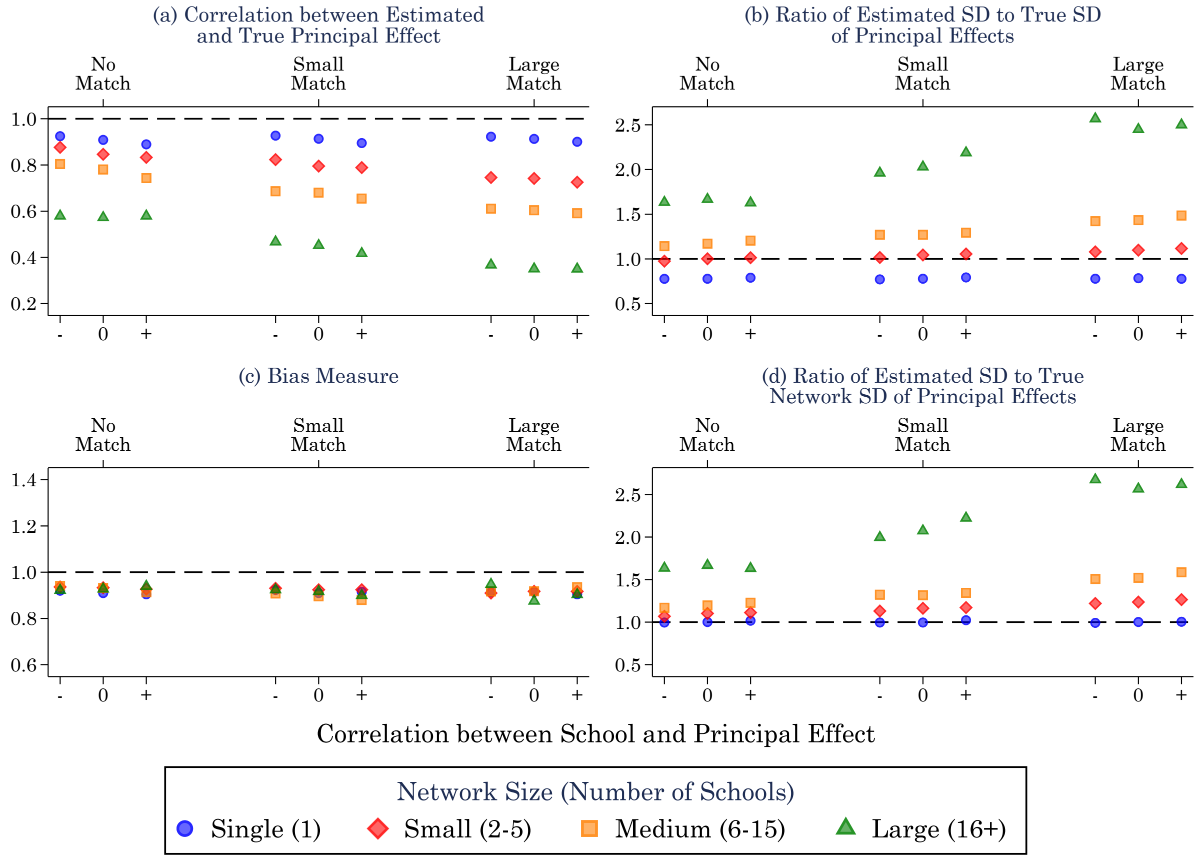


Figure 4: Principal FE + School FE Results by Network Size (Tennessee)

Notes: Results shown are only for the principal and school FE model in Tennessee. The legend defines the size of the connected network. In each plot, the y-axis is defined by the header. The bottom x-axis corresponds to the correlation between principal quality and the fixed school effect (-0.4, 0, 0.4). The top y-axis corresponds to the magnitude of the principal–school match effect (0%, 22%, 56% of the total principal effect). In the results shown here, the school performance effect constitutes 5% of the total variance in student test score growth. In Panels A, B, and C, the true principal effects are residualized on network fixed effects corresponding to the connected networks formed by principals and schools for the given VA measure.

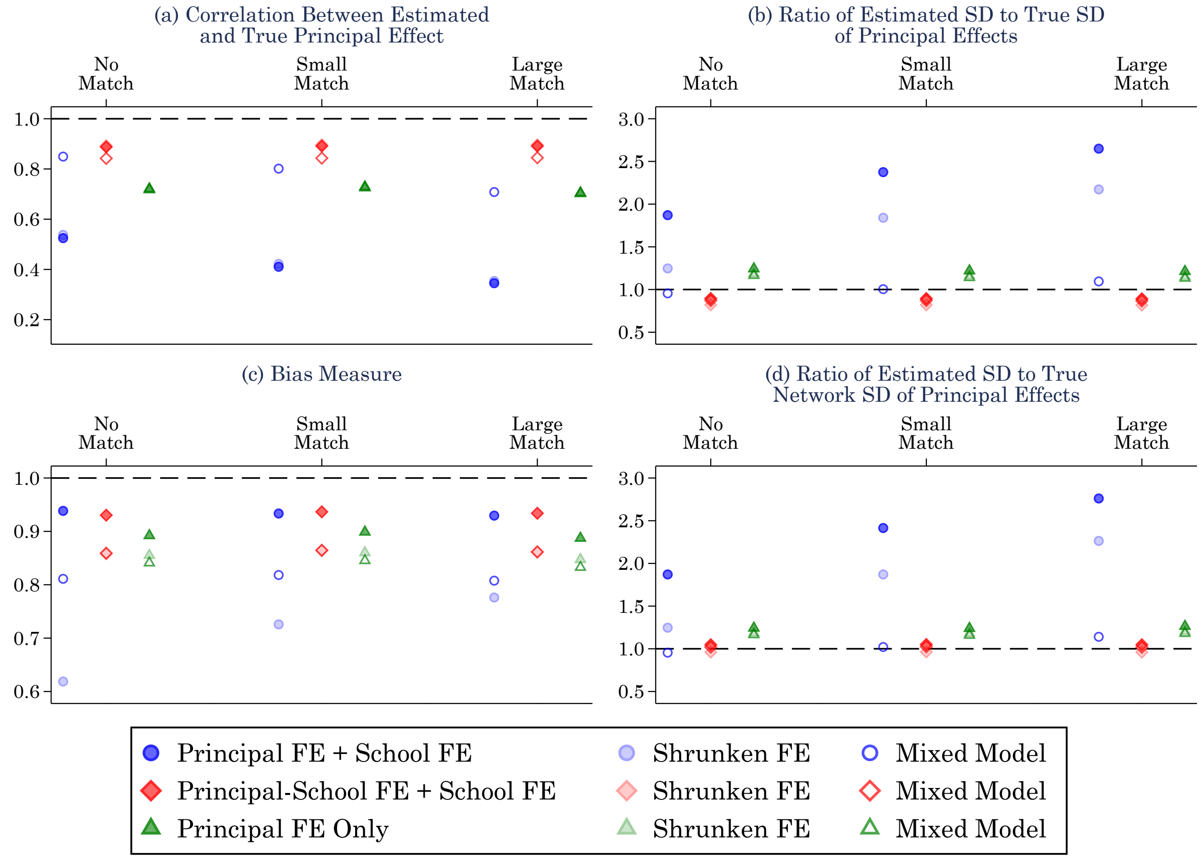


Figure 5: Comparison of Shrinkage Estimators

Notes: Shrinkage analysis only performed for principals in the largest connected network in Tennessee. In each plot, the y-axis is defined by the header. The top y-axis corresponds to the magnitude of the principal–school match effect (0%, 22%, 56% of the total principal effect). Additional horizontal spacing is to facilitate visual comparisons between models. In the results shown here, the school performance effect constitutes 5% of the total variance in student test score growth. In Panels A, B, and C, the true principal effects are residualized on network fixed effects corresponding to the connected networks formed by principals and schools for the given VA measure.

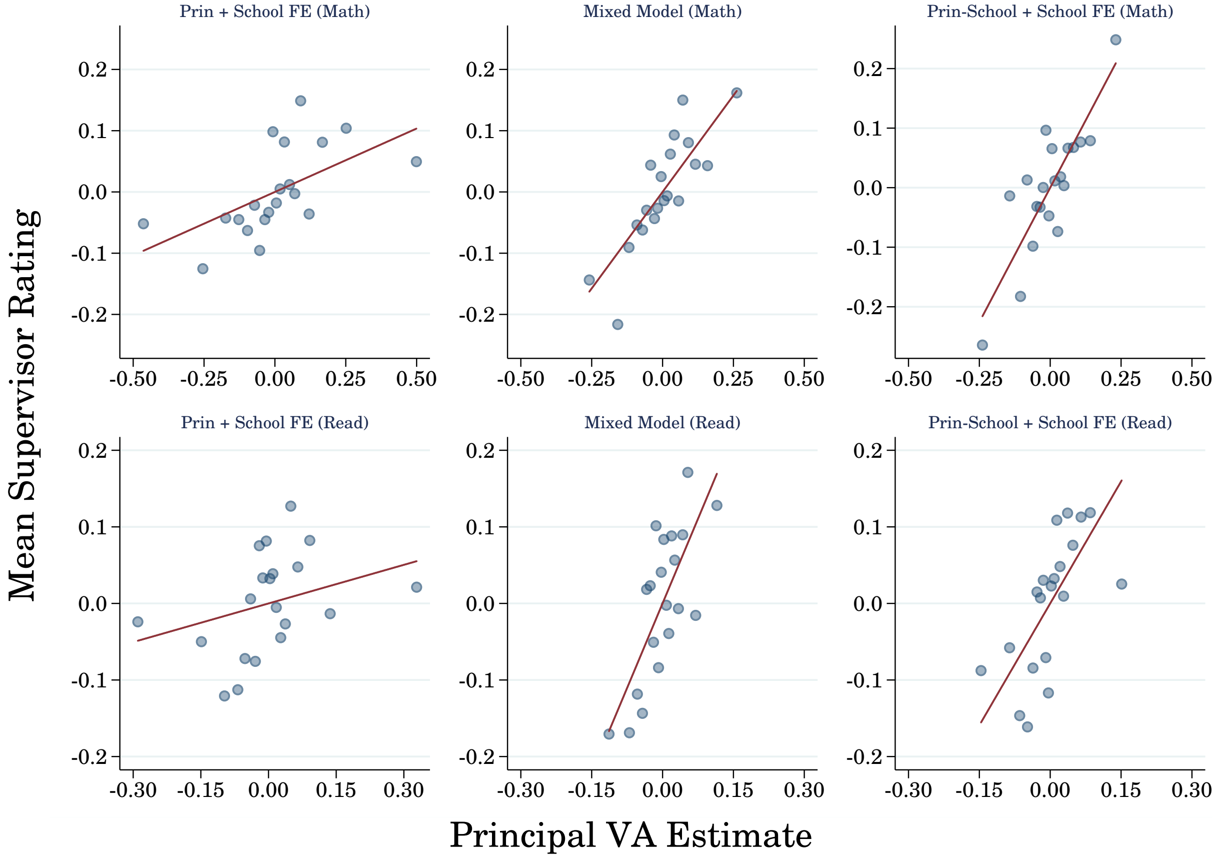


Figure 6: Relationship Between Principal VA Estimates and Supervisor Ratings in Tennessee

Notes: Each plot shows a binned scatterplot predicting supervisor ratings as a function of the principal VA estimate shown in the plot header, along with the OLS line. Models include fixed effects for the connected network corresponding to the principal VA estimate. The red line is estimated via OLS using the underlying data. Starting from the left, the standardized regression coefficient for the slope is (top row) 0.059, 0.105, 0.158; (bottom row) 0.031, 0.109, 0.116.

Table 1: Simulation Parameters

Panel A: Data-Generating Process

$A_{ijst} = \lambda A_{ijs,t-1} + \theta_{jst} + c_i + e_{ijst}$, where
$\theta_{jst} = \delta_{j(t,s)} + \alpha_{js(t,s)} + \gamma_s + v_{jst}$

Panel B: Parameters that are Fixed Across Simulations

λ (persistence)	0.5
$A_{ijs,t-n}$ (base score)	$Normal(0,1)$
c_i (fixed student effect)	$Normal(0,0.5)$
e_{ijst} (random deviation)	$Normal(0,1)$
Corr($A_{ijs,t-n}, c_i$)	0.5

Panel C: Parameters that Vary Across Simulations

School Performance Effect (θ_{jst}) % of Total Variance	Component Shares of School Performance Effect				Corr(δ_j, γ_s)
	δ_j	α_{js}	γ_s	v_{jst}	
5%	0.45	0.00	0.45	0.10	-0.4
10%	0.35	0.10	0.45	0.10	0
	0.20	0.25	0.45	0.10	0.4

Notes: This table summarizes the simulation parameters used to test the accuracy of principal value-added models. Panel A shows the data-generation process for student test scores. Panel B shows the parameters that are the same for every simulation. The base score ($A_{ijs,t-n}$) is drawn for each student's first observed year in the dataset, since there is no prior-year score. Panel C shows the parameters we vary across simulations. The component shares show the weight each component receives in the equation to produce the school performance effect, shown in the second line of Panel A.

Table 2: Network Statistics for Tennessee and New York Principals

	Network Size (# of Schools)				
	Single (1)	Small (2–5)	Medium (6–15)	Large (16+)	No Network
Panel A: Tennessee					
Number of Networks	564	174	22	2	
Mean Schools per Network	1	2.6	8.3	142.5	
Mean Principals per Network	2.5	5.8	20.5	353	
Number of Principals	1435	1003	450	706	243
Percentage of Total Principals	37.4	26.2	11.7	18.4	6.3
Mean Student Obs per Prin	1267	1457	1379	1470	2069
Connectivity (λ_2)	1.6664	0.6148	0.0438	0.0014	
Variance Inflation	0.002	0.004	0.007	0.027	
Panel B: New York City					
Number of Networks	695	117	12	2	
Mean Schools per Network	1	2.5	7.8	23	
Mean Principals per Network	2.8	6.5	19.6	56	
Number of Principals	1958	755	235	112	214
Percentage of Total Principals	59.8	23.1	7.2	3.4	6.5
Mean Student Obs per Prin	2837	2582	2330	2254	2362
Connectivity (λ_2)	1.5401	0.5050	0.0402	0.0051	
Variance Inflation	0.002	0.003	0.005	0.009	

Notes: Both connectivity (λ_2) and variance inflation are calculated following the approach of Jochmans and Weidner (2019). λ_2 is the smallest non-zero eigenvalue from the normalized Laplacian matrix that corresponds to each connected graph of principals. A smaller eigenvalue indicates that principals in a network are more weakly connected. Variance inflation is expressed in terms of the error variance of student test score growth. Principals without a network are those who were the only principal in their school across the study period.

Table 3: Standard Deviation of Empirical Principal VA Estimates in Tennessee

		Network Size (# of Schools)			
	All	Single (1)	Small (2–5)	Medium (6–15)	Large (16+)
Math					
Principal FE + School FE	0.201	0.094	0.152	0.259	0.323
Mixed Model	0.120	0.086	0.120	0.146	0.143
Principal-School FE + School FE	0.107	0.093	0.109	0.116	0.125
Reading					
Principal FE + School FE	0.127	0.054	0.083	0.195	0.203
Mixed Model	0.053	0.041	0.054	0.061	0.063
Principal-School FE + School FE	0.065	0.054	0.061	0.084	0.082

Notes: Table shows the standard deviation of principal VA estimates using actual student test scores. For the mixed model, the VA estimates are the best linear unbiased predictions (BLUPs) from a model with school fixed effects and principal random effects. The model-based estimate of the SD of the principal random effect is 0.140 in math and 0.067 in reading. The BLUPs are less variable due to shrinkage.

A Appendix Figures and Tables

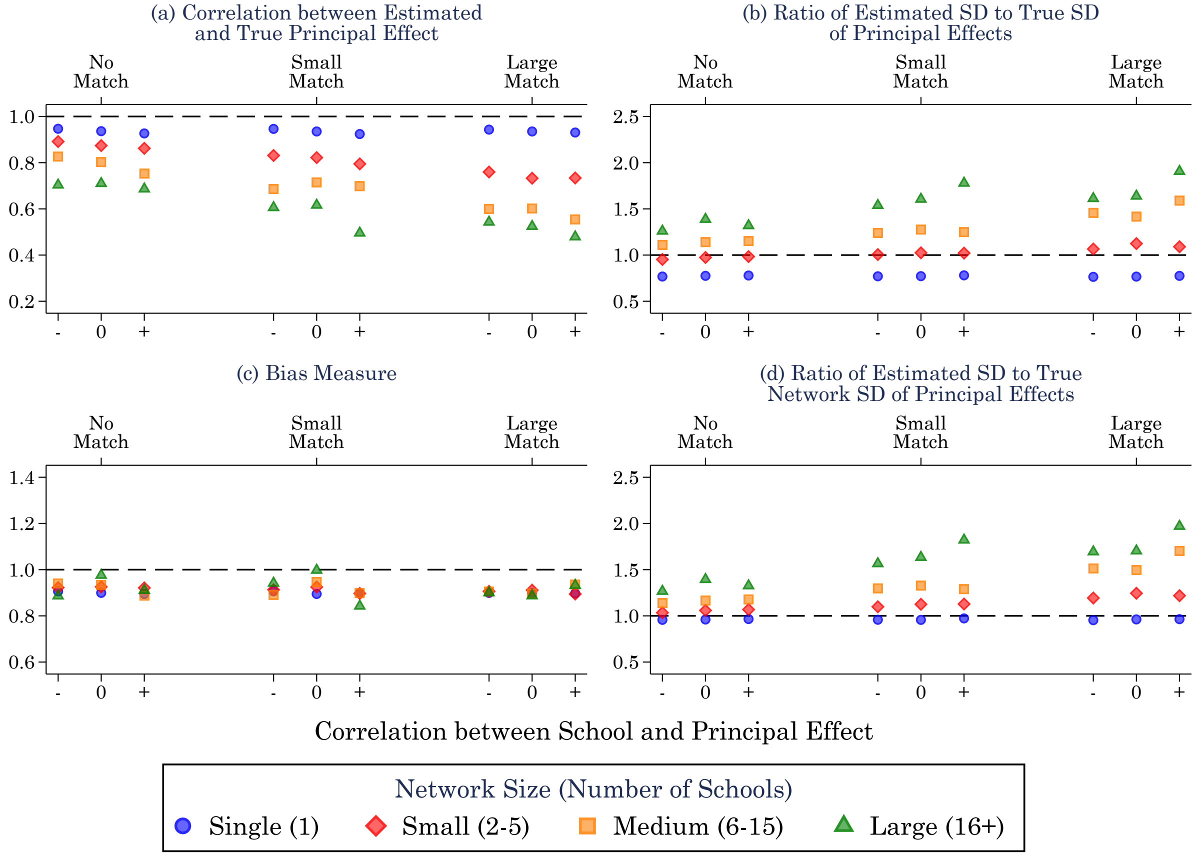


Figure A.1: Principal FE + School FE Results by Network Size (New York City)

Notes: Results shown are only for the principal and school FE model in New York City. The legend defines the size of the connected network. In each plot, the y-axis is defined by the header. The bottom x-axis corresponds to the correlation between principal quality and the fixed school effect (-0.4, 0, 0.4). The top y-axis corresponds to the magnitude of the principal–school match effect (0%, 22%, 56% of the total principal effect). In the results shown here, the school performance effect constitutes 5% of the total variance in student test score growth. In Panels A, B, and C, the true principal effects are residualized on network fixed effects corresponding to the connected networks formed by principals and schools for the given VA measure.

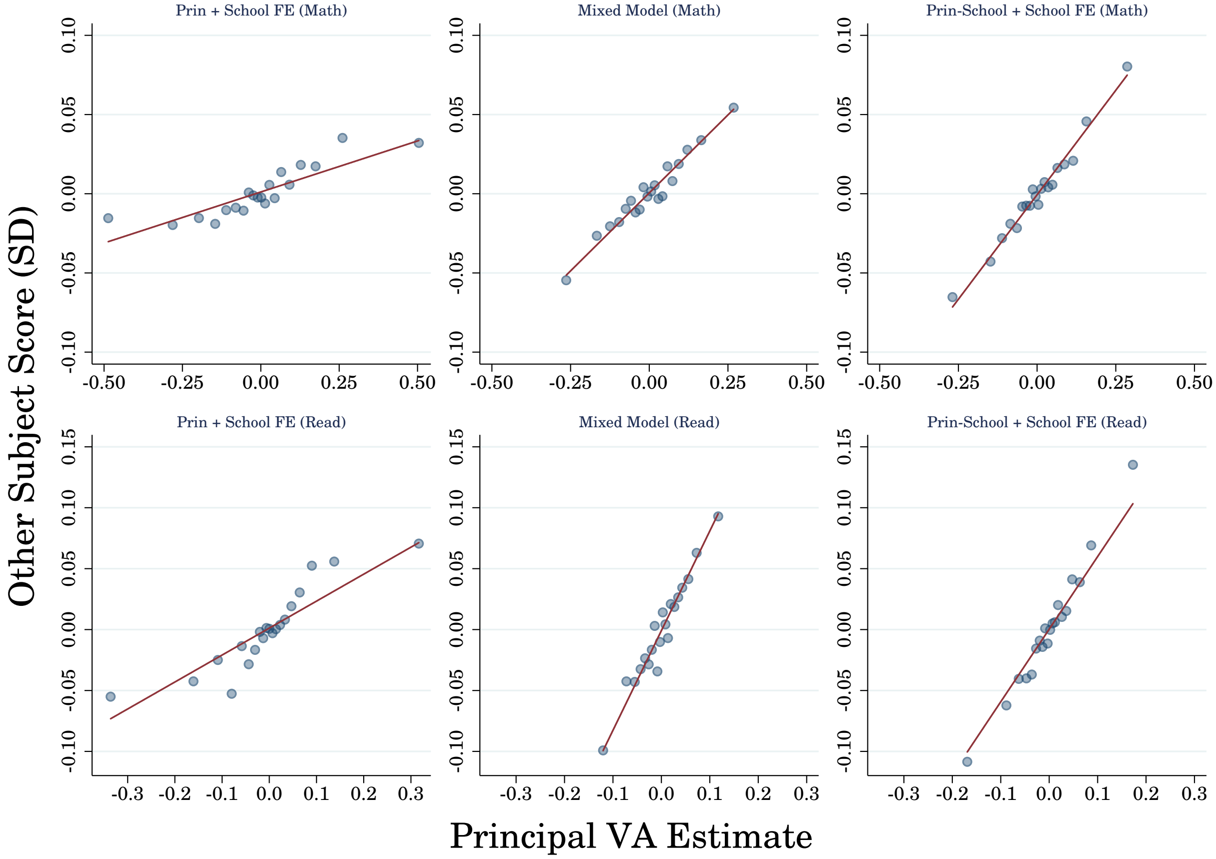


Figure A.2: Relationship Between Principal VA Estimates and Opposite-Subject Test Scores

Notes: Each plot shows a binned scatterplot predicting student test scores in the opposite subject as a function of the principal VA estimate for the subject shown in the plot header, along with the OLS line. Models include controls for students' prior-year test scores, demographic characteristics, school-by-year averages of these characteristics, fixed effects for the connected network corresponding to the principal VA estimate.

Table A.1: Standard Deviation of Empirical Principal VA Estimates in New York City

		Network Size (# of Schools)			
	All	Single (1)	Small (2–5)	Medium (6–15)	Large (16+)
Math					
Principal FE + School FE	0.070	0.044	0.088	0.144	0.146
Mixed Model	0.059	0.052	0.071	0.075	0.090
Principal-School FE + School FE	0.048	0.044	0.056	0.052	0.064
Reading					
Principal FE + School FE	0.051	0.033	0.065	0.098	0.143
Mixed Model	0.040	0.037	0.046	0.043	0.047
Principal-School FE + School FE	0.035	0.033	0.039	0.037	0.041

Notes: Table shows the standard deviation of principal VA estimates using actual student test scores. For the mixed model, the VA estimates are the best linear unbiased predictions (BLUPs) from a model with school fixed effects and principal random effects. The model-based estimate of the SD of the principal random effect is 0.082 in math and 0.057 in reading. The BLUPs are less variable due to shrinkage.

B Full Simulation Results

Table B.1: Tennessee, School Performance is 5% of Total Variance

θ_{jst} is **5%** of Total Variance	Corr, True Prin and Sch Eff	Corr, Est and True Prin Eff	Bias	SD Ratio	SD Ratio Within Net	Corr, Est Prin and Sch Eff
No Match Effect						
Principal FE + School FE	-0.4	0.76	0.93	1.10	1.23	-0.47
	0.0	0.74	0.92	1.12	1.26	-0.34
	0.4	0.73	0.92	1.12	1.26	-0.22
Principal-School FE + School FE	-0.4	0.92	0.93	0.83	1.02	-0.00
	0.0	0.90	0.92	0.84	1.03	0.00
	0.4	0.88	0.92	0.86	1.05	0.00
Principal-District FE + School FE	-0.4	0.87	0.93	0.94	1.08	-0.26
	0.0	0.83	0.92	0.97	1.11	-0.15
	0.4	0.81	0.92	0.99	1.14	-0.09
Principal FE Only	-0.4	0.58	0.60	1.03	1.03	.
	0.0	0.70	0.87	1.25	1.25	.
	0.4	0.83	1.19	1.44	1.44	.
Small Match Effect						
Principal FE + School FE	-0.4	0.67	0.92	1.21	1.38	-0.57
	0.0	0.64	0.91	1.25	1.42	-0.48
	0.4	0.62	0.91	1.31	1.49	-0.43
Principal-School FE + School FE	-0.4	0.92	0.93	0.83	1.02	-0.00
	0.0	0.90	0.93	0.84	1.03	0.00
	0.4	0.88	0.92	0.85	1.04	0.00
Principal-District FE + School FE	-0.4	0.78	0.92	1.02	1.18	-0.38
	0.0	0.77	0.92	1.03	1.19	-0.26
	0.4	0.75	0.91	1.05	1.21	-0.19
Principal FE Only	-0.4	0.60	0.63	1.04	1.05	.
	0.0	0.70	0.88	1.25	1.26	.
	0.4	0.81	1.15	1.41	1.42	.
Large Match Effect						
Principal FE + School FE	-0.4	0.56	0.92	1.46	1.68	-0.67
	0.0	0.56	0.91	1.42	1.64	-0.59
	0.4	0.55	0.91	1.45	1.68	-0.55
Principal-School FE + School FE	-0.4	0.91	0.93	0.84	1.02	-0.00
	0.0	0.90	0.93	0.84	1.03	-0.00
	0.4	0.89	0.93	0.85	1.04	-0.00
Principal-District FE + School FE	-0.4	0.71	0.92	1.11	1.30	-0.45
	0.0	0.71	0.91	1.10	1.29	-0.36
	0.4	0.69	0.91	1.12	1.32	-0.31
Principal FE Only	-0.4	0.62	0.69	1.09	1.11	.
	0.0	0.70	0.89	1.24	1.26	.
	0.4	0.79	1.10	1.36	1.39	.

Table B.2: Tennessee, School Performance is 10% of Total Variance

θ_{jst} is **10%** of Total Variance	Corr, True Prin and Sch Eff	Corr, Est and True Prin Eff	Bias	SD Ratio	SD Ratio Within Net	Corr, Est Prin and Sch Eff
No Match Effect						
Principal FE + School FE	-0.4	0.82	0.94	1.02	1.15	-0.39
	0.0	0.79	0.93	1.05	1.19	-0.26
	0.4	0.77	0.94	1.08	1.22	-0.20
Principal-School FE + School FE	-0.4	0.93	0.94	0.82	1.01	0.00
	0.0	0.92	0.93	0.83	1.02	0.00
	0.4	0.91	0.94	0.84	1.03	0.00
Principal-District FE + School FE	-0.4	0.88	0.93	0.93	1.06	-0.25
	0.0	0.87	0.94	0.94	1.08	-0.12
	0.4	0.85	0.94	0.96	1.11	-0.07
Principal FE Only	-0.4	0.59	0.60	1.02	1.02	.
	0.0	0.71	0.90	1.25	1.25	.
	0.4	0.84	1.20	1.43	1.43	.
Small Match Effect						
Principal FE + School FE	-0.4	0.68	0.94	1.22	1.39	-0.56
	0.0	0.67	0.93	1.22	1.39	-0.44
	0.4	0.67	0.94	1.23	1.40	-0.37
Principal-School FE + School FE	-0.4	0.93	0.94	0.83	1.01	0.00
	0.0	0.92	0.93	0.83	1.02	0.00
	0.4	0.91	0.93	0.84	1.02	0.00
Principal-District FE + School FE	-0.4	0.80	0.93	1.01	1.16	-0.36
	0.0	0.81	0.93	1.01	1.16	-0.23
	0.4	0.79	0.93	1.02	1.18	-0.17
Principal FE Only	-0.4	0.60	0.63	1.05	1.06	.
	0.0	0.71	0.89	1.24	1.25	.
	0.4	0.83	1.16	1.39	1.41	.
Large Match Effect						
Principal FE + School FE	-0.4	0.57	0.91	1.40	1.62	-0.66
	0.0	0.57	0.92	1.41	1.63	-0.60
	0.4	0.58	0.92	1.38	1.59	-0.51
Principal-School FE + School FE	-0.4	0.92	0.94	0.83	1.01	-0.00
	0.0	0.92	0.93	0.83	1.01	-0.00
	0.4	0.91	0.93	0.83	1.02	0.00
Principal-District FE + School FE	-0.4	0.72	0.92	1.10	1.28	-0.45
	0.0	0.72	0.92	1.09	1.27	-0.36
	0.4	0.71	0.91	1.10	1.28	-0.30
Principal FE Only	-0.4	0.62	0.69	1.08	1.10	.
	0.0	0.71	0.88	1.22	1.24	.
	0.4	0.80	1.10	1.36	1.38	.

Table B.3: New York City, School Performance is 5% of Total Variance

θ_{jst} is **5%** of Total Variance	Corr, True Prin and Sch Eff	Corr, Est and True Prin Eff	Bias	SD Ratio	SD Ratio Within Net	Corr, Est Prin and Sch Eff
No Match Effect						
Principal FE + School FE	-0.4	0.90	0.91	0.87	1.01	-0.20
	0.0	0.88	0.91	0.89	1.04	-0.10
	0.4	0.87	0.90	0.89	1.04	-0.04
Principal-School FE + School FE	-0.4	0.94	0.92	0.80	0.97	0.00
	0.0	0.93	0.91	0.81	0.98	-0.00
	0.4	0.92	0.90	0.81	0.99	-0.00
Principal FE Only	-0.4	0.56	0.57	1.01	1.01	.
	0.0	0.70	0.88	1.26	1.26	.
	0.4	0.83	1.21	1.46	1.46	.
Small Match Effect						
Principal FE + School FE	-0.4	0.85	0.91	0.91	1.07	-0.26
	0.0	0.84	0.91	0.93	1.09	-0.17
	0.4	0.81	0.90	0.94	1.11	-0.13
Principal-School FE + School FE	-0.4	0.94	0.92	0.80	0.98	0.00
	0.0	0.93	0.91	0.81	0.98	0.00
	0.4	0.92	0.91	0.81	0.99	0.00
Principal FE Only	-0.4	0.59	0.62	1.04	1.05	.
	0.0	0.70	0.88	1.26	1.27	.
	0.4	0.82	1.18	1.43	1.44	.
Large Match Effect						
Principal FE + School FE	-0.4	0.80	0.90	0.95	1.13	-0.32
	0.0	0.78	0.90	0.97	1.15	-0.25
	0.4	0.76	0.90	1.01	1.19	-0.23
Principal-School FE + School FE	-0.4	0.94	0.91	0.80	0.97	0.00
	0.0	0.93	0.91	0.80	0.98	0.00
	0.4	0.92	0.91	0.81	0.99	-0.00
Principal FE Only	-0.4	0.61	0.67	1.09	1.10	.
	0.0	0.70	0.88	1.26	1.27	.
	0.4	0.79	1.11	1.38	1.39	.

Table B.4: New York City, School Performance is 10% of Total Variance

θ_{jst} is **10%** of Total Variance	Corr, True Prin and Sch Eff	Corr, Est and True Prin Eff	Bias	SD Ratio	SD Ratio Within Net	Corr, Est Prin and Sch Eff
No Match Effect						
Principal FE + School FE	-0.4	0.91	0.91	0.86	1.00	-0.19
	0.0	0.91	0.91	0.86	1.00	-0.08
	0.4	0.89	0.91	0.87	1.02	-0.02
Principal-School FE + School FE	-0.4	0.95	0.92	0.79	0.97	-0.00
	0.0	0.94	0.91	0.79	0.97	0.00
	0.4	0.93	0.91	0.79	0.97	0.00
Principal FE Only	-0.4	0.57	0.58	1.00	1.00	.
	0.0	0.71	0.89	1.26	1.26	.
	0.4	0.84	1.23	1.46	1.46	.
Small Match Effect						
Principal FE + School FE	-0.4	0.86	0.92	0.91	1.07	-0.27
	0.0	0.85	0.92	0.92	1.08	-0.17
	0.4	0.84	0.90	0.92	1.07	-0.11
Principal-School FE + School FE	-0.4	0.94	0.92	0.80	0.97	0.00
	0.0	0.94	0.92	0.79	0.97	0.00
	0.4	0.93	0.91	0.80	0.97	0.00
Principal FE Only	-0.4	0.59	0.62	1.04	1.05	.
	0.0	0.70	0.89	1.26	1.26	.
	0.4	0.82	1.18	1.43	1.44	.
Large Match Effect						
Principal FE + School FE	-0.4	0.79	0.91	0.97	1.15	-0.33
	0.0	0.79	0.90	0.96	1.14	-0.24
	0.4	0.79	0.91	0.98	1.16	-0.21
Principal-School FE + School FE	-0.4	0.94	0.92	0.80	0.97	-0.00
	0.0	0.94	0.91	0.80	0.97	-0.00
	0.4	0.94	0.91	0.80	0.97	0.00
Principal FE Only	-0.4	0.62	0.68	1.09	1.10	.
	0.0	0.70	0.89	1.26	1.27	.
	0.4	0.80	1.12	1.39	1.40	.

C Variance Decomposition of Math Test Scores in Tennessee and New York City

To provide empirical support for our simulation parameters, we conducted variance decompositions for standardized math test scores in Tennessee and New York City. Specifically, we estimate via restricted maximum likelihood the following general form mixed model:

$$Y_{ijst} = \lambda(f(Y_{i,t-1})) + \gamma\mathbf{X}_{it} + \phi\mathbf{Z}_{st} + v_{ijst} \tag{1}$$

where v_{ijst} is a composite error term that includes either a (1) school-by-year random effect or (2) principal and school random effects. We control for prior-year test scores in both subjects and prior-year attendance rate (including squared and cubed terms), student characteristics (race/ethnicity, gender, FRPL-eligible, special education status, gifted status, flag for grade repetition, flag for within-year move to another school), and school-by-year means of the student characteristics.

Table C.1: Variance Components for Student Test Score Growth

	Tennessee				New York City			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A								
School-by-Year	18.6	9.8	9.6	8.5	27.0	8.0	6.5	5.4
Residual	81.4	90.2	90.4	91.5	73.0	92.0	93.5	94.6
Panel B								
Principal	4.1	5.8	5.9	6.0	3.4	2.2	2.3	2.2
School	18.0	6.1	5.7	6.2	24.0	4.6	3.0	2.0
Residual	77.9	88.2	88.5	87.8	72.6	93.2	94.7	95.8
Prior-year Test Scores		✓	✓	✓		✓	✓	✓
Student Characteristics			✓	✓			✓	✓
School Characteristics				✓				✓

Table C.1 shows variance components for the school-by-year random effects model in panel A and the principal and school random effects model in panel B. We show four specifications, beginning with an empty model and successively adding sets of controls. Panel A shows that the magnitude of the school-by-year random effect accounts for roughly 5–10% of the variation in residualized test scores. Panel B shows that even controlling for prior test scores, student characteristics, and school-by-year means of the student characteristics, there remains a substantial unobserved contribution of schools to the total variation in student test scores. Further, the magnitude of the principal and school random effects are roughly equal, underscoring the potential for bias in principal VA estimates that do not include school fixed effects.